基于混合策略的汉语最长名词短语识别

钱小飞[1] 侯敏[2]

（ 1.上海大学文学院 上海 200444；

2.中国传媒大学国家语言资源监测与研究中心有声媒体语言分中心 北京 100024）

摘要：提出一种基于语言知识评价的分类器集成方法，利用自动获得的搭配资源和人工评价规则，融合了基于支持向量机的最长名词短语识别结果和基于条件随机场的归约识别结果，进一步基于确定性规则有针对性地识别了分类器易出错的特殊结构，提高了对连续动词介词和连续名词造成的边界歧义的识别能力。实验取得了 89.30%的正确率和 89.62%的召回率，多词结构 F1 值较归约方法提高了 0.75%。

关键词：最长名词短语识别；语言知识评价；分类器集成；规则

Chinese Maximal Noun Phrase Recognition Based on Mixed Strategy

QIAN Xiao-fei[1], HOU Min[2]

(1. College of Liberal arts, Shanghai University, Shanghai 200444; 2. Broadcast Media Language Branch, National Language Resources Monitoring and Research Center, Communication University of China，Beijing 100024）

Abstract：This paper proposed a classifier ensemble method based on the language evaluation, and fused the MNP recognition results of SVMs and cascade CRFs based on reduction method, using the automatically obtained collocations and the manual assess rules. It then further targeted recognized the error-prone structures of the classifiers based on deterministic rules. The methods improve the recognition ability of boundary ambiguities of continuous verbs and prepositions as well as continuous nouns. The experiment is successful with a precision rate of 89.30% and a recall rate of 89.62%, especially it improves F1-score of multi-words MNPs by 0.75% in contrast with the reduction method.

Key words: Maximal Noun Phrase Recognition; language knowledge assess; classifier ensemble; rule

1 引言

最长名词短语（MNP）是句子中不被其他名词短语包含的名词短语，约占据句子长度的 60%以上，它的识别可以为完全句法分析以及机器翻译、指代消解等应用提供有效支持。

MNP 识别有三种方法：基于规则的方法[1]、基于统计的方法[2]和基于机器学习的方法[3]。其中，统计机器学习方法是当前的主流方法。从识别策略看，2-phase 策略[4]的和分类器集成的方法[5]取得了较好效果。以往研究关注算法改进，但对 MNP 的语言学特性关注不够，使识别系统过于依赖词（性）串等线性特征，导致复杂 MNP 和简单 MNP 识别 F1 值相差 13%-22%[5-6]；从识别策略看，2-phase 策略以较高训练代价提高识别精度，但也引入了级联错误[7]；而分类器集成方法多基于经验或数学手段获取基本分类器权重，系统复杂性因此大大提高，变得更加难以解释，分类对象的特点也很难得到充分的考虑。

针对以上问题，本文提出一种基于语言知识评价的分类器集成方法，融合非归约和归约的 MNP 识别结果，并基于确定性规则识别易出错的特殊结构，提高了 MNP 识别效果。

2 基于语言知识评价的集成

集成系统作出一个分类判断，并不一定以对象本身的运作规律为依据，我们难以知道，数学上的分类判断与识别对象本身的特点存在哪些必然联系。特别是当多数或全部基本分类器都做出错误分类时，没有机制能够提醒目标分类器，基本分类器作出了错误选择。

MNP 是一种复杂结构类型，涉及诸多歧义结构问题。本文的想法是针对分类器容易出错的具体类型，特别是一些典型歧义结构，利用更多的语言资源进行评价，得到一个基于语

言知识评价的集成系统。

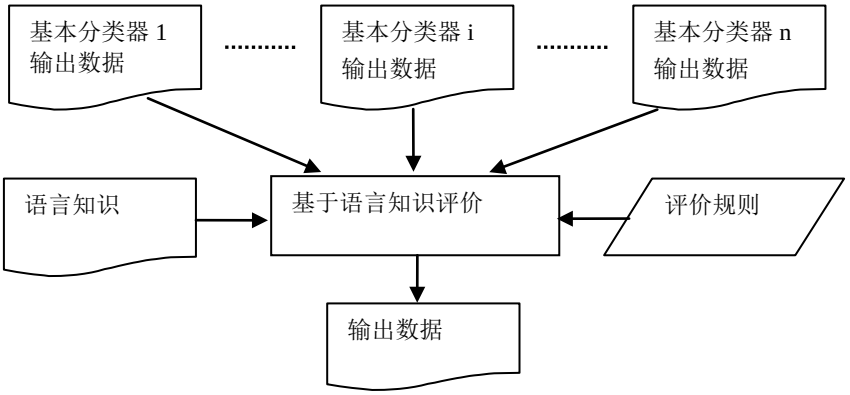

图1 基于语言知识评价的系统集成

尽管仍然要在基本分类器的分类结果之间作出选择，但不确定性已大大降低，我们有更确切的语言学证据说明，某些分类可能存在错误，应该如何进行选择，并且，在多数或全部基本分类器都发生错误的情况下，基于规则针对基本分类器的共同错误类型进行评价，仍然有可能取得正确的分类结果。

基于语言知识评价的集成方法把研究者的精力集中到语言知识获取和评价规则设计上，有针对性解决疑难问题，系统也具备了更好的可解释性；当然，语言知识和评价规则的分辨能力及覆盖率直接影响系统性能，因此，该方法与基于分类数据的评价方法各有优势。

具体来说，需要重点解决三个方面的问题：（1）如何获取语言知识，本文主要采用搭配知识；（2）如何设计评价规则；（3）如何基于语言知识进行评价。

2.1 搭配知识获取

2.1.1 搭配类型

词语搭配对识别复杂 MNP 非常有效，但由于词形数据稀疏和长距离依赖，在 ML 模型中难以充分利用。我们希望借助第三方资源，在分类器集成过程中发挥词语搭配的作用。针对 MNP 识别问题，选取动名搭配、介词搭配两种类型，五种搭配关系作为消歧资源。

动名搭配收录两种搭配关系：（1）动宾关系。确定动词宾语位置上MNP左边界。（2）定中关系。判定直接作定语的动词或状动结构，甄别错误的MNP左边界。

介词搭配收录三种搭配关系：（1）介宾搭配。判别介词宾语位置上MNP左边界。（2）介词框架。（3）介动搭配。后两者限定MNP范围，不可跨越介词搭配或介动搭配。

表 1 动名搭配示例

节点词	搭配词	搭配关系	搭配数据
跷	二郎腿	动宾关系	MI=19.90841
工作	人员	定中关系	Freq=87
凭	本事	介宾关系	MI=10.50395
为了	起见	介词框架	MI=9.203944
替	抱不平	介动关系	MI=12.79889

2.1.2 搭配获取

采用2004年《北京青年报》为语料，调用中科院计算所的ICTCLAS2009接口进行分词和词类标注，分类获取搭配。

（一）介词搭配获取

采用互信息方法获取关系较紧密的搭配实例和数据，并进行人工甄别。如果搭配统一表示为< *preItem*，*postItem* >，候选介词搭配自动获取方法如下：

(1) 扫描文本，计算词频 $f(w_i)$，语料总词数 *wordCount*

(2) 反向扫描文本

(a)如找到满足词类约束的 *postItem*，在小句范围内向前寻找 *preItem*，否则返回(2)

(b)如找到 *preItem*，计数搭配频率 $f(preItem, postItem)$

(3) 对每一组搭配< *preItem*，*postItem* >

(a) 计算词语出现概率 $P(preItem)$，$P(postItem)$，搭配概率 $P(preItem，postItem)$

(b) 计算互信息 $MI(preItem，postItem)$

(c)如果 $MI(preItem，postItem) > \alpha$，输出搭配对

（二）动词搭配获取

采用互信息和规则相结合的方法获取。为了尽可能获得长距离搭配，针对名词短语左递归的句法构造特征，设计了一个加权互信息获取模型（图2）。

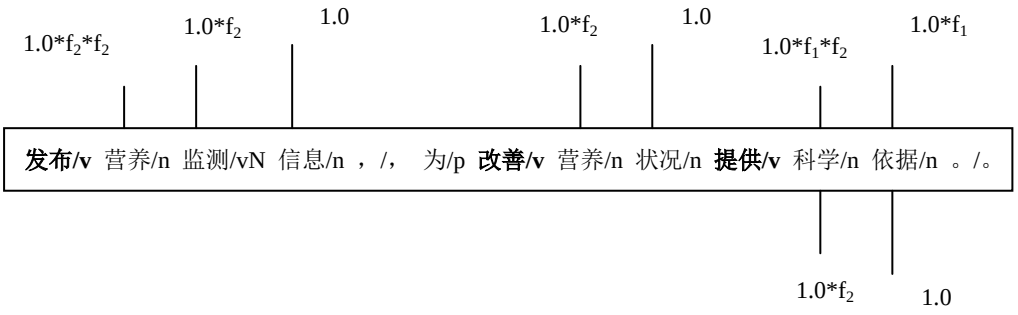

图2 加权互信息搭配获取方法示意图

以“改善/v”为例，在小句范围内分两步获取搭配：(1)识别一级候选搭配词“状况/n”，“依据/n”，初始化最近候选搭配词“状况/n”的频次为1，右部一级候选搭配词的频次依次乘以加权因子f_1，“依据”的权值为f_1；(2)以各一级候选搭配词为起点，反向扫描获取二级候选搭配词“科学/n”和“营养/n”，频次依次乘以加权因子f_2，分别为f_2和f_1* f_2。第(1)步如遇到结构助词“的”，重新以1为基础频次进行加权。

基于互信息的方法适用于高频实例，规则方法能获取稀疏搭配。搭配获取过程采用标注方式进行，保证句中搭配不被多条规则重复获取数据，分为两步：首先，标注动词边界有效性，识别“n|vN|aN+连词|语气词|标点|”等形式中的名词为中心名词N。其次，基于十条规则识别动宾搭配，推理定中搭配。动宾搭配规则包括：

（1） 对“v + 了|着|过 + x + [的 + x_1] + N”，如果x中不存在“的/u”、动词、介词等中断成分，识别“v N”为动宾搭配

（2）对“v v_1 N”结构，如果“v N”或“v_1 N”为动宾搭配，搭配频次加1

（3）对“[v_i x]$^+$ 的 N”结构，如果“v_i N”为动宾搭配，搭配频次加1

（4）对“v x 的 N”结构，如果x中不存在中断成分，识别“v N”为动宾搭配

（5）对“v x N”，如果x中不存在中断成分，识别“v N”为动宾搭配

定中搭配规则在获取动宾搭配后调用，包括：

（1）对“v 的 N”结构，识别“v N”为定中搭配

（2）对“的 + v_1 N”，识别“v_1 N”为定中搭配

（3）对“v v_1 N”结构，如果“v N”为动宾搭配，识别“v_1 N”为定中搭配

（4）对“v + x + [的] + v_1 N”，如果“v N”为动宾搭配，识别“v_1 N”为定中搭配

（5）对“v x v_1 x n_1 的 N”结构，如果“v N”为动宾搭配，识别“v_1 n_1”为定中搭配

基于规则获取的动宾搭配ruleVO，从基于统计获取的搭配中取得互信息数据。为减少错误搭配的干扰，采用ruleVO中互信息>5的搭配(16827条)进行实验，不进行任何人工干预。

2.2 评价规则设计

评价规则分为两种：词汇搭配规则、结构化规则。评价采用投票方式，分为得分和否决两种投票机制，否决票以-β（β=10）的形式表达。下文$word$表示单词，下标f表示MNP首词位置，下标h表示MNP尾词位置，$Score(MNP)$表示对MNP的评分。

2.2.1 词汇搭配规则

由两个部分组成：搭配信息和分值评价；主要有两个作用：评价当前MNP的可靠性；划分语块，评价当前MNP的合法性。

（1）边界有效性规则

边界有效性指动词或介词在当前语境下是否可以充当MNP左边界邻接词，包括两种类型。静态有效性由动词配价信息决定，如一价动词、能愿动词等一般不带名词性宾语；动态有效性由词语在句中的位置决定，如重叠式的第一个动词不能充当MNP左边界。

适用环境：$preItem$ MNP // 合同/d 作战/v [理论/n]

评价规则：如果$preItem$是无效边界，$Score(MNP) = Score(MNP) - \beta$

（2）框式搭配规则：介词框架、介动搭配等对应的规则

适用环境A：$preItem$ MNP $postItem$ // 在/p [结构/n 和/c 性能/n] 上/f

评价规则：如果$preItem$，$postItem$构成框式搭配，$Score(MNP) = Score(MNP) + 1$

适用环境B：$word_f$ ... $preItem$ ... $word_h$ … $postItem$

$preItem$ ... $word_f$ … $postItem$ … $word_h$

// 在/p [教师/n 工作/n 中/f 存在/v 的/u 一些/m 问题/n]

评价规则：如果$preItem$，$postItem$构成框式搭配，$word_f \ldots word_h$构成MNP，$Score(MNP) = Score(MNP) - \beta$

（3）交式搭配规则：动宾搭配、介宾搭配等对应的规则

适用环境A：$preItem$ $word_f$ ... $postItem_h$ // 付出/v [全部/n 心血/n]

评价规则：如果$preItem$，$postItem_h$构成交式搭配，$word_f \ldots postItem_h$构成MNP，$Score(MNP) = Score(MNP) + 1$

2.2.2 结构化规则

通过考虑歧义结构与MNP可能存在的位置关系，利用词汇搭配规则对不同的位置关系进行打分，优选最可能的位置关系。

令$head(phrase)$表示短语$phrase$的中心词；$syn(word_1, word_2)$表示$word_1$和$word_2$构成搭配，关系为syn；以MNP首词位置为0，$tag = tag^i$表示tag位于位置i，符号!表示否定。包括五组典型的结构或边界歧义模式，其中，bnp包含$baseNP$和单词块。

（1）p bnp De v．构成名词短语，如：对/p 敌人/n 的/u 仇恨/v；或者“bnp De v”位于介词框架内，如：在/p 他/rN 的/u 倡导/v 下/f 。评价规则如下：

当$p = p^0$，$vo(v, head(bnp))$，则$Score(MNP) = Score(MNP) + 1$

当$p = p^{-1}$，$vo(v, head(bnp))$，则$Score(MNP) = Score(MNP) - \beta$

（2）v bnp_1 De bnp_2．构成名词短语，如：没有/v 爱情/n 的/u 婚姻/n；或者作为动宾结构，如：揣摩/v 对方/n 的/u 心理/n 。评价规则如下：

当$v = v^0$，$vo(v, head(bnp_1))$，则$Score(MNP) = Score(MNP) + 1$

当$v = v^{-1}$，$vo(v, head(bnp_1))$，则$Score(MNP) = Score(MNP) - \beta$

（3）v n_1 n_2．构成名词短语，如：处理/v 问题/n 能力/n；或者作为动宾结构，如：保护/v 国家/n 财产/n 。评价规则如下：

当$v = v^0$，$!dz(v, n_1)$，$vo(v, n_2)$，则$Score(MNP) = Score(MNP) - \beta$

当$v = v^{-1}$，$dz(v, n_1)$，$!vo(v, n_2)$，则$Score(MNP) = Score(MNP) - \beta$

（4）v n．构成名词短语，如：作战/v 理论/n；或者动宾结构，如：度过/v 难关/n 。评价规则如下：

当$v = v^0$，$!dz(v,n)$，$vo(v,n)$，则$Score(MNP) = Score(MNP) - \beta$

当$v = v^{-1}$，$dz(v,n)$，$!vo(v,n)$，则$Score(MNP) = Score(MNP) - \beta$

（5）(v | p)$^+$ MNP．连续动词和介词分布造成的边界歧义，基于单一语料消歧，假设每个动词或介词都可作为左邻接词候选，调用前四条规则依次评价，取最优评价结果。

3 基于确定性规则的识别

分类器集成有利于发现识别错误，但两个基本分类器的相同错误难以发现和纠正，通常的办法是增加更多的基本分类器，不仅增加了系统复杂性，也缺少语言学依据。确定性规则针对易发生错误的结构类型，基于单个分类结果决断边界位置。主要处理六种情况：

（1）“的”字结构。当右边界为“的/u”时，向前寻找左邻接特征词“是、有、凡是、凡、像、如、为、特别是”等，如果找到，将左边界调整至左邻接特征词之后。

（2）“者”字结构。当右边界为“者/k”时，向前寻找左邻接特征词“凡是、凡”等，如果找到，将左边界调整至左邻接特征词之后。

（3）双宾结构。分类器常捆绑间接宾语和直接宾语。双宾结构规则对宾语重新划分。一些线性特征可作为判别依据，如出现双宾动词vSB，间接宾语是人称代词或称谓名词等。利用双宾动词词典$vSBDic$，间接宾语中心词词典N_1Dic，规则如下：

在“$vSB\ word_f\ word_{f+1} ... word_i\ word_{i+1}\ word_{i+2} ... word_h$”序列中，如果$word_i \in N_1Dic$，且$word_{i+1}$和$word_{i+2}$构成数量结构或指量结构，原序列调整为：$vSB\ word_f\ word_{f+1} ... word_h\ word_f\ word_{i+1} ... word_h$。

例如，“交给/vSB [我/rN 一/m 份/qN 材料/n]”，根据规则调整为，“交给/vSB [我/rN] [一/m 份/qN 材料/n]”。

（4）主谓谓语句。分类器常捆绑大主语和小主语。主谓谓语句规则重新划出大小主语。两者的语义距离可作为判别主谓谓语句的依据，距离大的相邻名词成分难以构成定中结构：A．人及其部分（HmPart），如“我 腿”；B．人及其心理（Mind），如“爸 爸 心情”；C．实体与实体，如“今天 中国”。短语在句中的位置影响主谓谓语句的判定，当“我 心情”位于句首主语位置时，常为大小主语关系，而处在宾语位置时，常作为名词短语。

主谓谓语句规则处理前两种情况，实体关系在相邻实体模块中处理。通过语义类别和语义关系判定语义距离，语义词典$semDic$记录词语语义类别，如条目“我/rN Human”等，关系词典$relDic$记录词语与词语、词语与语义类别的关系，如条目“@Human 心情/n Mind”，“*/nP 腿/n HmPrt”，其中@标识语义类别。规则表述如下：

在句首或者小句首位置，$word_f\ word_{f+1} ... word_i\ word_{i+1} ... word_h$序列中，如果$word_i$和$word_h$在3个词的窗口内满足$relDic$中的语义关系，则原序列重新划分为：$word_f\ word_{f+1} ... word_h\ word_f\ ... word_h$。

例如，“[群龙/nP 眼珠/n] 略微/dD 一/d 转/v”，根据规则调整为，“[群龙/nP] [眼珠/n] 略微/dD 一/d 转/v”。

（5）相邻实体。可以形成多种句法关系，如联合关系、修饰关系，也可以不形成句法关系，如大小主语，句子主语和状语等。调整规则分为合并规则和划分规则。

合并规则针对时间实体，如果基本分类器所识别的连续MNP中心词均为时间词，那么将连续的多个MNP合并为一个MNP。而划分规则处理三种情况：

A．人名|地名+时间短语。如果$word_f\ ... word_i\ word_{i+1} ... word_h$序列不包含动词和De，且$word_i$是人名，地名或处所词，$word_{i+1}$是时间词，或数词，$word_h$是时间词或时

间量词，则原序列重新划分为：$word_f$ ... $word_h$ $word_f$ ... $word_h$。

B．时间短语+指人代词序列。如果$word_f$ ... $word_i$ $word_{i+1}$ ... $word_h$序列不包含动词和De，且$word_i$是时间词，$word_{i+1}$不是时间词，$word_h$是指人代词，则原序列重新划分为：$word_f$ ... $word_h$ $word_f$ ... $word_h$。

C．时间短语+指人名词|地名|机构名。在时间词表中对时间词与指人名词、地名、机构名及普通名词的组合能力进行标注，并对不可修饰关系从严标注。

如果$word_f$ ... $word_i$ $word_{i+1}$ ... $word_h$序列不包含动词和De，且$word_i$是时间词，在以下四组条件下，原序列重新划分为$word_f$ ... $word_h$ $word_f$ ... $word_h$：

(a) 时间短语+人名。$word_i$不可修饰人名，$word_{i+1}$是人名、数词或代词，$word_h$是人名、指人代词或名词；(b) 时间短语+地名。$word_i$不可修饰地名，$word_{i+1}$是地名、数词或代词，$word_h$是地名、处所词或代词；(c) 时间短语+机构名。$word_i$不可修饰机构名，$word_{i+1}$是机构名、数词或代词，$word_h$是机构名；(d) 时间短语+普通名词短语。$word_i$不可修饰名词性成分，$word_{i+1}$是命名实体、数词或普通名词，$word_h$是命名实体或普通名词。

例如，“[过后/t 你们/rN 一/m 位/qN 同志/n] 找/v [我/rN] 要去/v 了/u”，根据规则(a)调整为“[过后/t] [你们/rN 一/m 位/qN 同志/n] 找/v [我/rN] 要去/v 了/u”。

（6）括号匹配

如果MNP跨越匹配括号的单个括号，则以右边界为基准，搜索第一个合法的动词介词或者中断标点（逗号等），作为左邻接词。如果右边界在右括号内，则左边界在括号内搜索；如果右边界在右括号外，则左边界从对应的左括号开始搜索。

4 系统流程

鉴萍[8]采用双向SVM作为基本分类器。我们的想法是，基于CRFs和基于SVMs的标注器，一个具有全局最优特性，一个具有确定性特点，其识别结果应该具有互补性；并且归约和非归约方法、正向标注和逆向标注策略也应该具有互补性。系统流程如下：

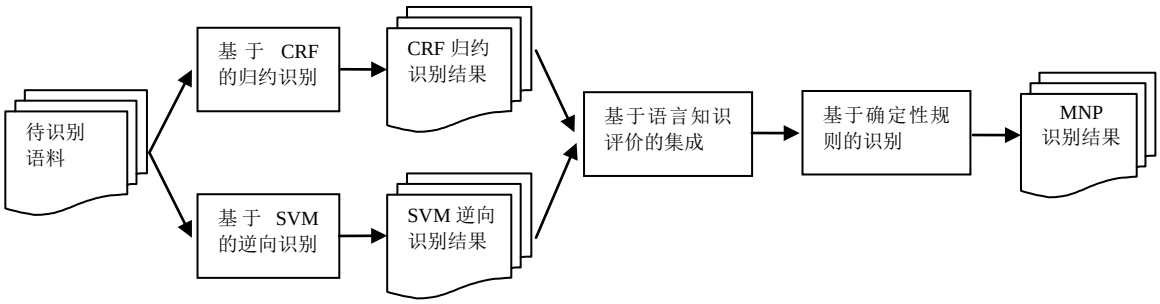

图3 系统流程图

集成系统采用两个MNP基本分类器。为了增加基本分类器的差异性，分别采用基于CRF模型的归约识别系统和基于SVM的非归约反向识别系统。前者基于2-phase策略，先识别baseNP，归约为中心词后识别MNP，并选择词形、词类、词长、义类（同义词词林三级义类）、baseNP核心为特征；后者优选了3元历史特征，词性和词类特征。集成系统和基于确定性规则识别的算法流程如下：

(1)标注动词、介词的边界有效性和介词搭配信息

(2)反向扫描两组分析结果：

(a)如果左右边界相同，评价连续的动词和介词

(b)如果右边界相同，左边界不同，评价左边界邻接词

(c)取评价最高的一个词语作为左邻接词，返回(2)

(3)正向扫描两组分析结果

(a)如果左边界相同，右边界不同，评价右边界邻接词

(b)取评价最高的一个词语右邻接词，返回(3)

(4)使用确定性规则识别，输出识别结果

其中，词语作为边界邻接词的评分首先使用词汇化规则和结构化规则，得分相同时使用词汇搭配数据（如互信息）评价。

5 实验结果及分析

对清华大学 TCT 树库进行 5 次无重复随机抽样，每个样本容量为 2000句。实验将每4个样本合并为训练语料，剩余 1 个样本作测试语料，构造 5 组训练测试对，进行交叉验证。5 组样本记作 sample5，每组训练测试对记为 samj，j∈[1,5]。

5.1 系统评测

采用正确率(prc)、召回率(rec)和调和平均值(F1-val)作为评价指标，基于混合策略的方法取得了89.30%正确率和89.62%的召回率。

表2 混合策略的实验结果

sum	num	prc	rec	F1
1	8268	89.65	89.85	89.75
2	8238	88.64	88.75	88.69
3	7970	89.38	90.33	89.85
4	8099	89.70	89.55	89.63
5	8103	89.13	89.63	89.38
ave	**–**	**89.30**	**89.62**	**89.46**

相比两个子系统，基于混合策略的方法较SVM逆向识别方法提高约2%，较CRF归约方法提高约0.5%。

表3 三种方法的比较

SVM 逆向扫描			CRF 归约			混合策略		
prc	rec	F1	prc	rec	F1	prc	rec	F1
87.27	87.50	87.39	88.68	89.21	88.95	89.30	89.62	89.46

由于融合规则和确定性规则更多地针对复杂MNP，专门考察多词结构的识别效果，新的系统相比CRF归约方法提高了0.75%左右。

表4 混合策略方法的实验结果（多词结构）

CRF 归约			混合策略		
prc	rec	F1	prc	rec	F1
84.59	85.09	84.84	85.40	85.79	85.60

尽管是小幅提高，但系统在每组样本上都有稳定的改善，并且所针对的需要实例搭配决策和特殊语言结构造成的问题正是统计识别的难点。

5.2 模块评测

以性能最优的基本分类器（CRF归约方法）为baseline，评价混合策略中的集成系统和确定性规则两组模块，每次递加一个模块进行测试。

表5　模块评测（F1-val）

模块	baseline	+集成系统	+确定性规则
测试	88.95	89.24	89.46

两组模块都能提高了识别效果，具体而言，集成系统和确定性规则分别改善了单宾语位置MNP和非宾语位置、双宾语位置MNP的识别效果。例如：

（1）“[不少/m 国家/n] 采取/v 促进/vJY【计算机/n 产业/n 兴旺发达/iV 的/u 政策/n]”修正为“[不少/m 国家/n] 采取/v 【促进/vJY 计算机/n 产业/n 兴旺发达/iV 的/u 政策/n]”

（2）“加强/v 对/p 【党员/n 的/u 思想/n 政治/n 教育/vN]” 修正为“加强/v 【对/p 党员/n 的/u 思想/n 政治/n 教育/vN]”

（3）“打/v 【补丁/n 的/u 裤子/n] 挽到/v [膝头/n]”修正为 “【打/v 补丁/n 的/u 裤子/n] 挽到/v [膝头/n]”

（4）“反正/d 没/v [人/n] 给/vSB [我/rN 一/m 分/qN 钱/n]”修正为“反正/d 没/v [人/n] 给/vSB [我/rN】【一/m 分/qN 钱/n]”

6 结语

本文提出了一种基于混合策略的MNP识别方法，包括基于语言知识评价的分类器集成和基于确定性规则的识别方法。前者利用自动获得语言学资源和人工总结的规则，融合了基于SVM逆向识别和基于CRF归约识别的结果；后者主要针对部分连续名词边界歧义问题，这是以往研究所没有关注到的，也是统计方法难以解决的问题。从识别难点看，仅使用少量自动获取的动宾搭配，使动词边界歧义和结构化歧义有所改善，但搭配覆盖率低、质量不够高是限制识别效果进一步提高的重要原因；此外，通过分析识别难点制定针对性的规则，如结构化规则(5)，部分解决了连续动词造成的边界歧义。进一步的工作包括提高动宾搭配的数量和质量，以及发掘更多的语言评价知识，如量名搭配知识，提高识别效果。

感谢清华大学周强老师为本文研究提供了 TCT 树库。

参考文献

[1] 周强,孙茂松,黄昌宁.汉语最长名词短语的自动识别[J].软件学报，2000,(2).

[2] 李文捷,周明,潘海华,等.基于语料库的中文最长名词短语的自动提取 [C]// 陈力为,袁琦.计算语言学进展与应用.北京：清华大学出版社,1995：119-124.

[3] 冯冲,陈肇雄,黄河燕,等.基于条件随机域的复杂最长名词短语识别[J].小型微型计算机系统,2006,(6).

[4] Chang-hao Yin. Identification of Maximal Noun Phrase in Chinese: Using the Head of Base Phrases [D]. POSTECH, Korea, 2005.

[5] Gui-ping Zhang, Wenjing Lang, Qiaoli Zhou, et al. Identification of Maximal-Length Noun Phrases Based on Maximal-Length Preposition Phrases in Chinese [C]// Proceedings of IALP 2010: 65-68.

[6] 代翠,周俏丽,蔡东风,等.统计和规则相结合的汉语最长名词短语自动识别[J].中文信息学报,2008,(6).

[7] Xue-Mei Bai, Jin-Ji Li, Dong-Il Kim, et al. Identification of Maximal-Length Noun Phrases Based on Expanded Chunks and Classified Punctuations in Chinese [C]// Proceedings of the 21st ICCPOL, 2006:268-276.

[8] 鉴萍,宗成庆.基于双向标注融合的汉语最长短语识别方法[J].智能系统学报,2009,(5).