**文章编号：**1003-0077（2011）00-0000-00

# CRFs 融合语义信息的英语功能名词短语识别*

马建军[1]，裴家欢[2]，黄德根[2]

（1. 大连理工大学外国语学院 辽宁 大连 116024；

2. 大连理工大学计算机科学与技术学院 辽宁 大连 116024）

**摘要：**名词短语识别在句法分析中有着重要的作用，而英汉机器翻译的瓶颈之一就是名词短语的歧义消解问题。研究英语功能名词短语的自动识别，则将名词短语的结构消歧问题转化成名词短语的识别问题。基于名词短语在小句中的语法功能来确定名词短语的边界，选择商务领域语料，采用了细化词性标注集和条件随机域模型结合语义信息的方法，识别了名词短语的边界和句法功能。在预处理基于宾州树库细化了词性标注集，条件随机域模型中加入语义特征主要用来识别状语类的名词短语。实验结果表明，结合金标准词性实验的 F 值达到了 89.04%，改进词性标注集有助于提高名词短语的识别，比使用宾州树库标注集提高了 2.21%。将功能名词短语识别信息应用到 NiuTrans 统计机器翻译系统，英汉翻译质量略有提高。

**关键词：**功能名词短语；名词短语识别；条件随机域模型；语义信息

**中图分类号：**TP391　　**文献标识码：**A

# Identification of English Functional Noun Phrases Using CRFs Combining the Semantic Information

Ma Jianjun [1], Pei Jiahuan [2], Huang Degen [2]

(1. School of Foreign Languages, Dalian University of Technology, Dalian Liaoning 116024, China ;

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

**Abstract:** Noun phrase identification plays an important role in parsing and one major problem with English-Chinese machine translation lies in its ability to resolve the ambiguous problems caused by nouns. Therefore, a study on the automatic identification of a kind of English functional noun phrases (NP) may transform the task of resolving structural ambiguity caused by noun phrases into the task of NP chunking. Functional noun phrases refer to those noun phrases which are defined based on their syntactic functions in clauses. Made on a corpus of business domain, this study aims to identify both the scope of NP chunks and their syntactic function types, by refining the Part-of-speech (POS) tagset, and adopting conditional random fields (CRFs) model combining the semantic information. Modification is made based on the Penn Treebank tagset in the pre-processing, and semantic features are added to the CRFs model to help improve the recognition of the adjunct types of noun phrases. Test results show that the system has achieved an F-score of 89.04% in the open test using our gold standard tags, and refining the POS tagset is a better approach for NP chunking, which has increased the F-score by 2.21%, compared with the model using the Penn Tree bank POS tags. This knowledge of English functional noun phrases is then combined with the NiuTrans SMT system, which slightly improves the English Chinese translation performance.

**Key words:** functional noun phrases; noun phrase identification; CRFs; semantic information

---

* **收稿日期：**　　**定稿日期：**

**基金项目：** 教育部人文社会科学研究规划基金项目（13YJAZH062）

**作者简介：**马建军（1972——），女，教授，主要研究方向为句法分析、机器翻译；裴家欢（1992——），女，博士在读，主要研究方向为句法分析、查询时间意图分类和句子相似度计算；黄德根（1965——），男，教授，主要研究方向为自然语言处理、机器翻译。

## 1 引言

名词短语识别在句法分析中有着重要的作用，名词短语的识别可以降低句法分析的复杂性，提高机器翻译的性能和效率。英汉机器翻译的瓶颈之一就是名词短语的歧义消解问题，真实文本中存在的大量名词短语结构歧义是导致整句英汉机器翻译正确率较低的主要因素之一。人工翻译中看似简单的名词短语结构往往却在机器翻译中产生结构歧义。比如说：n1+prep+n2 结构。结合名词短语在小句中的句法功能，这一表层结构至少存在 3 种深层结构如下：

a. He likes the book on the table. （prep+n2 结构“on the table”作后置定语）
b. He finds the book on the table. （prep+n2 结构“on the table”作状语）
c. He puts the book on the table.（prep “on”是小品词，put…on…是固定搭配）

将句子输入到 GOOGLE 在线翻译系统，得到如下结果：

a. 他**喜欢的书放在桌子**上。（参考译文：他喜欢**桌子上的书**。）
b. 他发现**在桌子上的书**。（参考译文：他**在桌子**上找到了那本书。）
c. 他**把**书**放在桌子上**。（参考译文：他**把**书**放在桌子上**。）

从前两个例句可以看出，在统计机器翻译中，往往把 prep+n2 简单处理为 n1 的后置定语，造成明显的翻译错误。因此专门针对机器翻译领域，研究英语名词短语的结构歧义及消歧方法，对提高机器翻译的效率，将起到关键作用。

目前的英语名词短语识别研究主要集中在基本名词短语和最长名词短语的识别。Church[1]利用统计方法进行名词短语的识别，Voutlainen[2]设计了名词短语识别系统 NPTool，但是这两个系统识别的名词短语非常简单，甚至不包括名词前的修饰成分。Ramshaw 和 Marcus[3]提出了基本名词短语的概念，把名词之前的修饰语包含在名词短语中。Koehn 和 Knight[4]提出了最长名词短语的定义，把名词后的修饰语包含在名词短语中。这两种名词短语是根据名词短语的逻辑结构来定义的，如：是否包括名词前和名词后的修饰语，而没有考虑名词短语的句法功能。文献[5]研究发现，这种定义方法在识别阶段易于识别，但是在翻译阶段会引起许多结构歧义。因此有必要融合结构和句法功能来定义名词短语，把对翻译要素的考虑提前到句法分析阶段，提高句法结构歧义的消歧率和机器翻译的质量。马建军和黄德根[6]基于系统功能语法获取名词短语在小句中的句法功能，根据句法功能界定名词短语的边界，并将这种名词短语定义为功能名词短语，初步论证了这种界定方法在机器翻译应用中的实际意义。

国内外英语名词短语的识别方法有很多，主要可分为两大类：基于规则的方法和基于统计的方法。基于规则的方法主要指通过人工方法或人工结合机器学习的半自动方法获取规则，如：基于转换的错误驱动学习方法[7]。基于统计机器学习的方法包括：边界统计方法[1]，基于实例的方法[8]，基于粗糙集的方法[9]，基于决策树[10]，基于词频统计模型[11]，以及支持向量机方法[12-13]。从统计模型的角度看，主要有最大熵模型[14-16]，隐马尔可夫模型[17-18]，条件随机域模型[19-20]等。研究的趋势是综合多种不同的方法以及应用不同的统计模型来识别名词短语，如规则和边界统计相结合[21]，最大熵和规则方法相结合[22]，基于条件随机域和支持向量机的混合统计模型[23-24]。

因此，本文选择商务领域语料，采用了细化词性标注集和条件随机域模型结合语义信息的方法，进行功能名词短语的自动识别研究，不仅识别名词短语的边界，同时还识别名词短语的句法功能。

## 2 英语功能名词短语的定义

本文识别的功能名词短语是指由中心名词及其修饰语组成的短语。其结构为 “前置修饰语+名词+后置修饰语”。其中，前置修饰语可以是限定词、数词、形容词、或名词；名词包括普通名词、代词和专有名词；后置修饰语可以是介词或“介词+名词短语”结构或形容

词；前置修饰语和后置修饰语不是必须的结构。基于系统功能语法[25]，本文把名词短语在小句中的功能主要归纳为 6 类：S，C，C1/C2/C3/C4，D，PR，和 CR。其含义如表 1 所示。

表 1　名词短语功能块标注集

| 功能块类型 | 中文含义 | 英文含义 |
|---|---|---|
| S | 主语 | subject |
| C | 补语 | complement |
| C1/C2/C3/C4 | 补语 1/补语 2/ 补语 3/补语 4 | the first/second/ third/fourth complement |
| D | 状语 | adjunct |
| PR | 谓语的一部分 | the residues of process |
| CR | 补语的一部分 | the residues of complement |

具体例句如下：

a. [S A very clever traveling salesman] sold [C his complete stock of washing machines] [D the next day].

b. Please send [C1 us] [C2 all available data on your Hand Tools], enabling [C us] to introduce [C1 your products] to [C2 our customers].

c. If [S your products] are satisfactory and [S prices] are right, [S we] expect to place [PR regular orders for] [C large numbers].

## 3 标注训练语料

### 3．1 人工标注训练语料

本文所用的是自建的小型商务英语语料库。由 10059 个经过去重的英语句子及其中文翻译构成，包含 14 个类别，如：询价及回复、运输、建立业务、还价、合同、包装、运输、付款、代理、索赔、订货、保险、报价和市场营销。根据功能块标注集对近 20 万词的英语语料进行了人工标注，语料库的语料信息如表 2 所示。

表 2　语料库信息

| 信息 | 数目 |
|---|---|
| 句子 | 10,059 |
| 词性标记 | 198,127 |
| 名词短语标记 | 42,319 |
| 名词短语的功能块类型 | 9 |

表 3 详细列举了名词短语功能块的分布情况。表 3 表明，语料中名词短语的句法功能归纳为 9 个：S，C，D，C1，C2，C3，C4，PR，CR。其中，主语（S），补语（C），和状语（D）是名词短语在小句中的三个主要句法功能，一共占整个语料的近 84%；而 C3，C4 和 CR 则出现频率很小，一共才占 0.23%。值得注意的是，状语占 15.83%，是名词短语识别的重点，因为不像主语和补语，状语往往包括那些诸如“for your reference”之类的以介词开头的名词短语，而不是以名词开头的名词短语。这些名词短语以介词开头，在识别中很容易被误认为是之前名词的后置定语，因而造成识别错误，对机器翻译带来结构歧义问题。

表 3　名词短语功能块的分布

| 名词短语功能块类型 | 数目 | 比率 (%) |
|---|---|---|
| S | 15,625 | 36.92 |
| C | 13,115 | 30.99 |
| C1 | 2,622 | 6.20 |
| C2 | 2,622 | 6.20 |
| C3 | 53 | 0.12 |
| C4 | 5 | 0.01 |
| D | 6,698 | 15.83 |
| PR | 1,537 | 3.63 |
| CR | 42 | 0.10 |
| 所有 | 42,319 | 100 |

### 3．2 IOB2 标注方法

在实验中，将名词短语的识别任务转化为序列标注任务。采用 IOB2 的标注方法，对名词短语块的边界进行标记，从而把块分析问题转化为序列标记问题。标记 B 表示当前词是名词短语的首词，标记 I 和标记 O 分别表示当前词属于名词短语内还是名词短语外。同时，标记 I 和标记 B 还同名词短语的句法功能结合起来，如：B-S 表示当前词是名词短语的开始，该名词短语的句法功能是 S（主语）。具体范例见表 4。

表 4　IOB2 标注方法范例

| 词 | 词性标记 | IOB2 标记 |
|---|---|---|
| A | DT | B-S |
| very | RB | I-S |
| clever | JJ | I-S |
| travelling | VBG | I-S |
| salesman | NN | I-S |
| sold | VBD | O |
| his | PRP$ | B-C |
| complete | JJ | I-C |
| stock | NN | I-C |
| of | INP | I-C |
| washing | NN | I-C |
| machines | NNS | I-C |
| the | DT | B-D |
| next | JJ | I-D |
| day | NN | I-D |
| . | . | O |

## 4 研究方法

### 4．1 CRFs 的识别模型

本文将功能名词短语的识别问题转化为序列标注问题，利用条件随机域建立功能名词短语的序列标注模型。本文介绍的条件随机域模型是比较简单的线性链条件随机域，给定参数 $\Lambda=(\lambda_1,\lambda_2,\ldots,\lambda_n)$，线性链条件随机域定义在一个给定的观测序列 $X=(x_1,x_2,\ldots,x_n)$ 上对应的状态标记序列 $Y=(y_1,y_2,\ldots,y_n)$ 的条件概率为：

$$P_\Lambda(Y|X)=\frac{1}{Z_X}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},X,t)\right) \quad (1)$$

其中 $Z_X$ 是所有状态序列的归一化因子，使得在给定输入上所有可能状态序列的概率之和为 1。$f_k(y_t,y_{t-1},X,t)$ 通常被定义为关于整个观测序列和位置 $t$ 以及位置 $t-1$ 标记的二值特征向量函数，参数 $\lambda_k$ 是在训练中得到的与特征函数 $f_k$ 相关的权重，当训练状态序列被完全明确地标记后，可为该模型找到最优的 $\lambda$ 值，一旦这些值被找到，一个新的未标记序列的标记工作就可以用 Viterbi 算法来完成，k 的取值范围取决于模版中特征的数量。那么求解序列标注的任务就是求出使条件概率 $P_\Lambda(Y|X)$ 最大的 $Y$，即最大可能的标记序列为：

$$Y^*=\arg\max P_\Lambda(Y|X) \quad (2)$$

条件随机域模型识别名词短语的关键在于特征的选择，特征的选择恰当与否会对识别结果产生直接的影响。通常来讲，丰富的上下文特征对于识别精确率的提高有着积极的作用。

本文在进行特征选择的时候，不仅充分利用了词和词性本身的信息，考虑到词和词性及其上下文之间存在着的种种依赖关系，还利用了融入更多上下文信息的组合特征。在实验中，本文选择了三种主要特征：当前词，当前词的词性，以及组合特征。表 5 为条件随机域模型所采用的特征模板，其中 $w_i$ 代表词本身特征，$t_i$ 代表词的词性特征，其他特征为词和词性的组合特征。特征模板描述如下：

（1）前后各三个词的词语和词性特征。

（2）相邻两个词的词性组合特征。

（3）次相邻两个词的词性组合特征。

（4）当前词的词性分别与前、后词的词语组合特征。

（5）相邻两个词的词性组合特征再分别与其正对应窗口为四的词语组合特征。

（6）后两个词的词性组合特征再分别与当前词、前词的词性组合特征。

其中，最后两条特征是通过大量的特征选择实验总结得出的对结果有较大影响的特征组合。利用表 5 中的特征模板，将给定的训练语料拿到 CRFs 上进行训练，再用训练得到的名词短语识别模型对测试语料进行标注，最后得到功能名词短语的识别结果。

表 5　条件随机域模型的特征模板

| 编号 | 特征 |
|---|---|
| 1 | $w_i$； $t_i, i \in [-2,2]$ |
| 2 | $t_i t_{i+1}, i \in [-1,0]$ |
| 3 | $t_i t_{i+2}, i \in [-1,0]$ |
| 4 | $w_{-1} t_0$； $w_1 t_0$ |
| 5 | $t_{-1} t_0 w_i, i \in [-2,1]$ |
| 6 | $t_1 t_2 t_i, i \in [-1,0]$ |

## 4．2 CRFs 结合语义信息

通过大量的语言现象可以发现，一些“介词+名词”搭配的短语对于提高功能块的标注效果有积极的作用，如 for your reference 为标注整个句子的组块标记提供了重要的信息。为了进一步利用这种固定搭配短语的特征，本文进一步引入语义信息，即用语义类来代替固定搭配中的名词部分，这样一定程度上减少了数据稀疏的影响（具体见表 6）。

表 6　带语义信息的功能名词短语标注举例

| 词 | 词性 | 语义标记 | IBO2 标记 |
|---|---|---|---|
| We | PRP | 0 | B-S |
| are | VBP | 0 | 0 |
| attaching | VBG | 0 | 0 |
| relevant | JJ | 0 | B-C |
| specifications | NNS | To3 | I-C |
| for | INP | 0 | B-D |
| your | PRP$ | 0 | I-D |
| reference | NN | For1 | I-D |
| . | . | 0 | 0 |

本文的语义信息是从词典《柯林斯 COBUILD 英语语法句型 2：名词与形容词》[26]中人工抽取形成的。每个名词只赋予一个语义标记，若超过一个，则选择出现频率最高的情况，另外对于不在词典中的词则统一用数字 0 来标识。表 7 以“for + N”搭配为例说明了语义分析的结果。当引入语义后，一些低频的搭配短语可以聚集在一起。例如：for your reference（供您参考）和 for your consideration（供您决定）可以分类为“for + N”的搭配中，同时，reference 和 consideration 根据语义还可以进一步分类为“USE”类。这样 reference 和 consideration 就属于相同的语义类：For/USE 组，并赋予相同的语义标记：For1。

表 7　名词语义信息列表

| 标记 | 语义类名 | 名词 |
|---|---|---|
| For1 | For/use 组 | correction, consideration, file, information, reference, record,… |
| For2 | For/benefit 组 | benefit, convenience, purpose,… |
| For3 | For/time 组 | being, moment, time, years,… |
| … | … | … |

此时，在 CRFs 模型中引入固定搭配特征。通过观察可以发现，一个搭配短语对于标记短语中的每一词的 BIO 状态，以及前词后词的 BIO 状态有重要的提示作用。因此，为了捕获这些搭配短语的信息，需要在位置 t 位于搭配短语中时，启动特征抽取过程。对于前词和后词的位置并不需要这样，因为它们的标记可以通过$f_k(y_t, y_{t-1}, X, t)$来影响其标注。依旧以表 6 为例，抽取到的特征如下所示：

当 t 指向 for 时：　$s_1 = For1$

当 t 指向 your 时：　$s_2 = For1$

当 t 指向 reference 时：　$s_3 = For1$

其中，s 的下标指示当前词在短语中的位置，等式右边为短语的语义标记。

利用结合语义信息的 CRFs 模型进行了预实验，主要关注功能块状语 D 的识别。将英文语料按照商务情景分成 7 组，每组任意抽取 300 句，共 2,100 句子作为测试语料，其余 7,959 句子作为训练语料。评价指标包括功能名词短语的准确率（Precision，*P*）、召回率（Recall，*R*）和 F 值（F-1 measure，$F_{\beta=1}$）。具体计算公式如下：

$$P = \frac{\text{识别出的正确名词短语数}}{\text{识别出的名词短语数}} \quad (3)$$

$$R = \frac{\text{识别出的正确名词短语数}}{\text{原文本中的所有的名词短语数}} \quad (4)$$

$$F_{\beta=1} = \frac{2 \times P \times R}{P+R} \times 100\% \quad (5)$$

功能块 D 识别结果如表 8 所示。结果表明，加入语义信息后，识别结果有所提高，准确率、召回率和 F 值分别提高了 3.76%，4.56%和 4.16%。

表 8　功能块 D 标注结果

| 模型 | *P* (%) | *R* (%) | *F* (%) |
|---|---|---|---|
| Baseline | 77.14 | 76.04 | 76.59 |
| Baseline+语义信息 | 80.9 | 80.6 | 80.75 |

## 4．3 预处理

预处理进行了词性标注。为了提高名词短语识别效果，本文面向机器翻译的目的在宾州树库词性标注集[27]的基础上构建了本文的词性标注集。具体改进方法如表 9 所示，主要在四个方面进行了细化，包括：区分介词和从属连词；增加了功能词 it, for, by，如：It/IT is

dangerous for/FOR children to walk alone in the forest 中的 it 和 for；区分单词 to 的不同功能；定义小品词的广义定义，即与动词构成短语动词的介词或方位副词，包括如：He informed Barbara of/RP his objections.中的 of，而这个小品词在宾州树库中标注为介词 IN。

表 9　两个标注集的比较

| | 本文的标注集 | | 宾州树库标注集 | |
|---|---|---|---|---|
| 从属连词或介词 | INC | 从属连词 | IN | 从属连词/介词 |
| | INP | 介词 | | |
| 功能词 | IT, FOR, BY | | 无 | |
| 单词 “to” | TO | 不定式 | TO | |
| | INP | 介词 | | |
| | RP | 小品词 | | |
| 小品词 | RP | 广义小品词 | RP | 狭义小品词 |

## 5 实验结果及分析

### 5．1 封闭测试和开放测试实验结果

应用 CRFs 结合语义信息和规则的方法进行了封闭测试和开放测试，开放测试采用 5 重交叉验证方法，分别进行了结合金标准词性标记（gold standard POS tags）和结合实际输出的词性标记两种实验。为检验本文的词性标注集在功能名词短语识别中的作用，在开放测试中还选择了斯坦福标注器的词性标记来取代本文的词性标记，分别进行了上述两种相同的实验。实验结果如表 10 所示。

表 10　名词短语识别结果

| 实验 | 测试类型 | 词性标记 | P(%) | R(%) | F(%) |
|---|---|---|---|---|---|
| 实验 1 | 封闭 | 本文金标准 | 99.56 | 99.54 | 99.55 |
| 实验 2 | 开放 | 本文金标准 | 89.47 | 88.62 | 89.04 |
| 实验 3 | 开放 | 斯坦福金标准 | 87.52 | 86.15 | 86.83 |
| 实验 4 | 开放 | 本文实际输出 | 86.95 | 84.86 | 85.89 |
| 实验 5 | 开放 | 斯坦福实际输出 | 84.89 | 82.63 | 83.74 |

实验 1 的结果表明，封闭测试识别的准确率达到 99.56%，召回率达到 99.54%，F 值达到 99.55%。实验 2 和实验 3 结合金标准词性标记进行了 5 重交叉实验，分别基于本文的词性标注集和宾州树库词性标注集，基于宾州树库标注集的金标准词性标记是通过斯坦福词性标注器标注后，人工修订标注结果得到的。根据表 10，使用本文的词性标注集的识别结果要好于使用宾州树库词性标注集。使用本文的金标准词性标记准确率达 89.47%，召回率 88.62%，F 值达到 89.04%，这个结果比使用宾州树库词性标注集的结果分别提高了 1.95%，2.47%和 2.21%。实验 4 和实验 5 结合实际输出的词性进行了实验。将实验 4 和实验 5 的结果分别同实验 2 和实验 3 的结果进行比较，结果表明，无论是采用本文的词性标注集，还是采用宾州树库词性标注集，使用实际输出的词性标记的识别结果低于使用金标准词性结果，准确率低近 2.5 个百分点，召回率和 F 值的差值都超过了 3 个百分点。这说明，需要提高词性标注器的标注效果，从而为名词短语的识别提供更好的支持。另外，同实验 2 和实验 3 的结果一样，采用本文的实际词性标记的识别结果仍然高于采用斯坦福词性标记的识别结果，这也说明了选择词性标注集对名词短语的识别有一定的影响。

### 5．2 六种功能块识别结果

表 11 比较了结合金标准词性的 2 个试验中（实验 2 和实验 3），六种名词短语功能块 S，C，D，PR，C1，C2 的识别结果，识别结果用平均值表示。没有比较功能块 C3，C4 和 CR 的识别结果，是因为名词短语以这三种功能块出现的频率较小，在语料中分别占 0.12%，0.01%和 0.10%（见表 3）。从表 11 可以看出，几乎在所有六种功能块的识别中，使用本文的词性

标注集的识别结果都好于使用宾州树库标注集。仅有一种情况除外，即结合斯坦福金标准词性的 PR 的召回率（74.93%）略高于结合本文的金标准词性的召回率（74.75%）。但是在其他所有情况，无论是准确率还是召回率和 F 值，都是基于本文的词性标注集的结果好。此外，表 11 还表明在结合本文的金标准词性标记的实验中，S，C，C1 的识别结果要比 D，PR，C2 好得多。S 的识别结果最好，F 值达到 97.46%；而 D 的识别仍然是研究的难点，F 值为 79.47%。所以，状语 D 的识别问题值得进一步研究。

表 11　功能块识别结果

| 类型 | 本文词性标注集 | | | 宾州树库标注集 | | |
|---|---|---|---|---|---|---|
| | P (%) | R (%) | F (%) | P (%) | R (%) | F(%) |
| S | 97.83 | 97.09 | 97.46 | 97.43 | 96.59 | 97.01 |
| C | 83.80 | 86.40 | 85.08 | 82.17 | 85.11 | 83.61 |
| D | 80.05 | 78.92 | 79.47 | 77.45 | 76.22 | 76.82 |
| PR | 86.51 | 74.75 | 80.19 | 86.08 | 74.93 | 80.10 |
| C1 | 91.71 | 85.72 | 88.55 | 89.03 | 79.09 | 83.73 |
| C2 | 83.70 | 78.04 | 80.67 | 81.03 | 71.35 | 75.83 |

### 5．3 在统计机器翻译中的应用

将功能名词短语信息应用到 NiuTrans 统计机器翻译系统中，以检验功能名词短语识别对机器翻译质量的影响。随机选择 2000 英汉句对作为测试语料，其余 8059 句对作为训练语料，应用 NiuTrans 统计机器翻译系统构建英汉机器翻译 baseline；然后将英语功能名词短语的句法信息作为特征加入到生成的短语表。比较两次翻译的 BLEU 值，结果见表 12。结果表明：翻译结果略有提高，BLEU 值从 9.87%提高到 10.42%，提高了 0.55%。

表 12　统计机器翻译结果

| 模型 | BLEU 值 （%） |
|---|---|
| Baseline | 9.87 |
| Baseline+短语信息 | 10.42 |

## 6 结论

本文改进了词性标注集，采用了 CRFs 结合语义信息的方法识别英语功能名词短语。实验结果表明：

（1）使用 CRFs 结合语义信息的方法能有效识别英语功能名词短语，使用本文的金标准词性标记准确率达 89.47%，召回率 88.62%，F 值达到 89.04%。

（2）细化词性标注集有助于提高功能名词短语的识别。结合金标准词性标记的开放测试结果表明，使用细化的词性标注集比使用宾州树库标注集 F 值提高了 2.21%。结合实际输出的词性标记的开放测试也表明，采用细化的实际词性的识别结果仍然高于采用斯坦福词性的识别结果，F 值提高了 2.15%。

（3）功能名词短语识别的主要问题集中在作状语的名词短语识别方面。

（4）在统计机器翻译系统中加入功能名词短语识别信息，略微提高了英汉机器翻译的质量，BLEU 值提高了 0.55%。

功能名词短语识别可以应用到机器翻译的研究中，因为识别这类名词短语能够在识别阶段就解决了名词短语结构歧义问题，把名词短语的结构消歧问题转化成名词短语的识别问题。如果这类名词短语在识别阶段能够较好地识别出来，就能够在一定程度上提高机器翻译的质量。

## 参考文献


[1] Church K. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text[C]// Walker D.E. Proceedings of Second Conference on Applied Natural Language Processing. Austin, USA: Association for Computational Linguistics, 1988: 136-143.

[2] Voutilamen A. NPTool, A Detector of English Noun Phrases[C]// Church K. W. Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives. Columbus, USA: Association for Computational Linguistics, 1993: 48-57.

[3] Ramshaw L, Marcus R. Text Chunking using Transformation-Based Learning[C]// Ejerhed E. Proceedings of the Fourth Workshop on Very Large Corpus. Copenhagen, Denmark: Association for Computational Linguistics, 1995: 82-94.

[4] Koehn P, Knight K. Feature-Rich Statistical Translation of Noun Phrases[C]// Hinrichs E. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics, 2003: 311-318.

[5] 马建军. 基于规则和统计的机器翻译方法歧义问题比较分析[J].大连理工大学学报(社会科学版), 2010, 31(3): 114-119.

[6] 马建军,黄德根.英语功能名词短语的研究及其应用[J].大连理工大学学报(自然科学版), 2012, 52(1):126-131.

[7] Brill E. Transformation-based error-driven parsing[C]// Bunt H. Proceedings of the Third International Workshop on Parsing Technologies. Tiburg, Netherlands: Association for Computational Linguistics, 1993:13-16.

[8] Veenstra, J.,Buchholz S. Fast NP Chunking Using Memory-Based Learning Techniques[C]// Verdenius F. Proceedings of the Eighth Belgian-Dutch Conference on Machine Learning. Wageningen, Netherlands: Wageningen ATO-DLO, 1998:71-78.

[9] 郭永辉,杨红卫,马芳,等. 基于粗糙集的基本名词短语识别[J]. 中文信息学报, 2006, 20(3):14-21.

[10] 李生, 孟遥. 基于决策树的英语 BNP 识别[J]. 黑龙江工程学院学报, 2001, 15(1):36-39.

[11] Kong, L., Ren, F., Sun, X. et al. Word Frequency Statistics Model for Chinese Base Noun Phrase Identification[C]// Huang D.S. Proceedings of the 10th International Conference on Intelligent Computing (ICIC). Taiyuan, China: Springer International Publishing, 2014:635-644.

[12] Kudo, T., Magsumoto,Y.Chunking with support vector machines[C]// Levin L. Proceedings of NAACL-2001. Pittsburgh, USA: Association for Computational Linguistics, 2001: 192-199.

[13] Wu, Y. C., Lee Y. S., Yang J. C. Robust and Efficient Multiclass SVM Models for Phrase Pattern Recognition[J]. Pattern Recognition, 2008(41): 2874–2889.

[14] Koeling, R. 2000. Chunking with Maximum Entropy Models[C]// Cardie C. Proceedings of CoNLL-2000 and LLL-2000. Lisbon, Portugal: Association for Computational Linguistics, 2000:139-141.

[15] 周雅倩, 郭以昆, 黄萱菁,等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算

机研究与发展, 2003, 40(3): 440-446.
[16] 王晓娟, 赵春. 最大熵方法在英语名词短语识别中的应用研究[J]. 计算机仿真, 2011, 28(3): 414 -417.
[17] Molina, A., Pla F. Shallow Parsing using Specialized HMMs[J]. Journal of Machine Learning Research, 2002(2): 595–613.
[18] Shen, H., Sarkar, A. Voting between Multiple Data Representations for Text Chunking[C]// Kégl B. Proceedings of the Eighteenth Meeting of the Canadian Society for Computational Intelligence, Canadian AI. Victoria, Canada: Springer Berlin Heidelberg, 2005:389-400.
[19] Sha, F. Pereira, F. Shallow Parsing with Conditional Random Fields[C]// Hearst M. Proceedings of HLT-NAACL 2003. Edmonton, Canada: Association for Computational Linguistics, 2003:213-220.
[20] Sun, X., Morency, L. P., Okanohara, D. et al. Modeling Latent-Dynamic in Shallow Parsing: A Latent Conditional Model with Improved Inference[C]// Scott D. Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, UK: Association for Computational Linguistics, 2008:841-848.
[21] 梁颖红,赵铁军,翟舒. 规则和边界统计相结合的英语基本名词短语识别[C]// 孙茂松. 全国第七届计算语言学联合学术会议论文集. 哈尔滨, 中国：中文信息学会，2003: 173-178.
[22] 吕琳,刘玉树. 最大熵和 Brill 方法结合识别英语 BaseNP[J]. 北京理工大学学报, 2006, 26(6): 500-503.
[23] 谭魏璇, 孔芳, 倪吉,等. 基于混合统计模型的中文基本名词短语识别[J]. 计算机应用与软件, 2011, 28(8): 254-156.
[24] 钱小飞, 侯敏. 基于混合策略的汉语最长名词短语识别[J]. 中文信息学报, 2013, 27(6): 16-22.
[25] Halliday M A K. 功能语法导论[M]. 北京: 外语教学语研究出版社, 2008.
[26] Sinclair J. 柯林斯 COBUILD 英语语法句型 2：名词与形容词[M].上海: 上海外语教育出版社, 2000.
[27] Marcus, M. P., Santorini, B., Marcinkiewicz, M. A. Building a large annotated corpus of English: the Penn Treebank[J]. Computational Linguistics, 1993, 19(2):313–330.

作者联系方式：
1. 马建军 地址：大连市甘井子区凌工路 2 号，大连理工大学外国语学院文科楼；邮编：116024；电话：13941175918；电子邮箱: majian@dlut.edu.cn

2. 裴家欢 地址：大连市甘井子区凌工路 2 号，大连理工大学创新园大厦 A0933；邮编：116024；电话：18741124605；电子邮箱: p_sunrise@mail.dlut.edu.cn

3. 黄德根 地址：大连市甘井子区凌工路 2 号，大连理工大学创新园大厦 A0933；邮编：116024；电话：15804251073；电子邮箱: huangdg@dlut.edu.cn