自动词性标注中语法因素和词汇因素对英汉语的不同影响*

邢富坤[1]，宋 柔[2]
[1]解放军外国语学院，河南 洛阳 471003
[2]北京语言大学 语言信息处理研究所，北京 100083
E-mail: xingfukun@tom.com; songrou@blcu.edu.cn

摘 要：本文使用词性自动标注模型对影响英汉语词性标注的相关因素进行定量研究，进而探究词汇因素与语法因素各自对英汉语词性标注的影响，目的是为深入分析英汉语在词类问题上的差别，更好地构建汉语语料库提供参考依据。本文将词汇因素近似地形式化为词汇发射概率和词汇最大词性概率，语法因素近似地形式化为词性转移概率，并利用隐马尔科夫模型、马尔科夫模型和词汇最大概率模型进行自动标注实验。通过比较不同模型的标注准确率，发现语法因素对于英语词性标注的影响显著大于汉语，词汇因素对汉语词性标注的影响显著大于英语。在英汉语内部，英语的词的核心语义因素与句法因素对词性标注的影响基本相仿；而汉语的词的核心语义因素对词性标注的影响显著高于句法因素。

关键词：词类；英汉对比；隐马尔科夫模型；马尔科夫模型；词汇最大概率模型

Study on Grammatical and Lexical Factors in Automatic PoS Tagging Between English and Chinese

Xing Fukun[1], Song Rou[2]
[1] PLA University of Foreign Languages, Luoyang 471003
[2] Center for Language Information Processing, Beijing Language and Culture University, Beijing 100083
E-mail: xingfukun@tom.com; songrou@blcu.edu.cn

Abstract: This article makes a quantitive study on the basis of automatic PoS (Part of Speech) tagging model to explore the different roles of lexical and grammatical factors in English and Chinese PoS tagging. The main goal of this article is to make an in-depth study on the differences in the issue of word classes between English and Chinese and give some suggestions to the construction of Chinese corpus. The lexical factor is formalized as lexical emission probability and the grammatical factor is formalized as the PoS transition probability. Hidden Markov Model, Markov Model and Lexical Maximum Probability Model are used to do the PoS tagging experiment. The experiment results show that the grammatical factor in English plays a more important role than that in Chinese. The lexical factor in Chinese has more significant effect on PoS tagging than that in English. In English lexical semantic factor and syntactic factor equally affect PoS tagging. In Chinese lexical semantic factor is more important than syntactic factor in PoS tagging.

Keywords: parts of speech, comparative study between English and Chinese; HMM; MM; MP

1 词类问题概述

词类是语言学研究中的重要范畴，词类研究构成了语言学研究的重要基础。在语言工程领域，词类研究也处于十分重要的地位。但汉语词类研究还很不完善，尤其是当前词类研究不仅要面向语言教学，更要面向机器的语言自动处理，由于机器对于语言知识内在逻辑性的严格要求以及实际应用任务的严格检验，都使得汉语现有词类体系和词类知识暴露出诸多问题，这些问题已经引起研究者的普遍关注，例如宋柔（2009，2011）中指出了面向语言工程的汉语词类体系面临的本质性困难。

对于词类划分的依据，目前大部分语言学界的研究者都认为英汉语都以语法功能为依据划分

* 本文的工作得到国家自然科学基金（60872121）的资助。

词类（朱德熙，1983；俞士汶，1998），但是还是存在不同的观点。郭锐先生区分划类的依据和划类的标准，并指出“词类划分的依据是词的内在表述功能或词的语法意义”（郭锐，2002）。一些学者从宏观层面对汉语和其他语言做出了对比分析，最具代表性的是现代语言学之父洪堡特在1826年《论汉语的语法结构》一文中对于汉语语法特点的论述：

> 汉语语法最根本的特性我认为是在于这样一点，即，汉语不是根据语法范畴来确定词与词的联系，其语法并非基于词的分类；在汉语里，思想联系是以另一种方式来表达的。（姚小平译 2001）

潘文国先生（1997）也指出，英汉语在词法与句法上存在显著差异，“中国模仿西洋语法，煞费苦心地建筑起来的语法大厦，其基本的矛盾是词法和句法合不拢，……，这样的分类就失去了意义。究其原因，是因为词分类的标准与句成分的设立标准不一。”

前人有关英汉语词类特点的结论大多是从理论层面给出的，近年来更多学者从人类语言类型的大背景下对汉语词类展开研究，其中Bisang（2008）对古汉语的词类研究具有较强代表性。

综合考察前人研究成果，可以发现绝大部分有关英汉词类对比的研究都是建立在语言学家自身的语感经验之上，这种研究方法下得出的结论一方面具有较强的抽象性和概括性，但另一方面却由于缺少在较大规模真实语料上的验证，因此不同观点之间会产生争论，且由于缺少客观的评价标准，争论很难得以评判和解决。

本文以较大规模的英语和汉语词性标注语料库为基础，从自动词性标注的角度对英汉词类问题进行比较，考察不同因素对英汉语词性标注的影响，从而探究英汉语在词类问题上更深层差异，回答影响英汉语词性标注的主要因素是否相同，不同因素对英汉语的词性标注存在何种影响等问题，以期能够加深对汉语词类问题的认识，探索符合汉语特点的词属性描写方法和语料库加工内容。

本文的基本结构是，第一节对英汉语词类问题进行概述，第二节介绍与词性标注相关的因素，第三节介绍词性自动标注模型，第四节介绍实验设计，第五节实验结果报告，第六节是对全文的小结。

2 自动词性标注及相关因素

2.1 自动词性标注

自动词性标注是机器利用已有的知识和特定的模型算法，为文本中的每一个词标注上合适词性的过程（Manning，2005）。自动词性标注可以使用基于规则的方法，也可以使用基于统计的方法。衡量自动标注结果的指标主要是总体标注准确率，其计算方法是：

$$\text{总体标注准确率} = \frac{\text{正确标注的词例数}}{\text{总词例数}}$$

衡量自动标注结果的另外两个主要指标分别是兼类词的标注准确率和未登录词的标注准确率。本文除了考察总体标注准确率外，还考察了兼类词标注准确率。兼类词是指在标注词典或训练语料中具有不止一种词性的词，其计算方法是：

$$\text{兼类词标注准确率} = \frac{\text{正确标注的兼类词词例数}}{\text{兼类词的总词例数}}$$

2.2 影响自动词性标注的相关因素

影响自动标注准确率的主要因素可以分为语言因素和非语言因素两大类。

语言因素可以分为：①句法因素；②词法因素；③词的核心语义因素。①和②构成影响词性标注的语法因素，②和③构成影响词性标注的词汇因素。需要注意的是②兼属于语法因素与词汇

因素，而由于汉语基本上不存在词的形态变化，因此②对汉语词性标注的影响基本可以忽略不计；但由于在英语等语言中存在较为丰富的形态变化，形态变化对词性标注有着重要的指示作用，因此②对于英语的词性自动标注有着较为重要的影响。

非语言因素主要包括训练与测试的语料规模和人工标注质量等。

3 词性标注模型

词性标注任务可以描述为：给定词序列 W=$w_0w_2...w_h$，求该序列对应的概率最大的词性序列 $\hat{Q}=q_0,...,q_h$。下面就本文中使用的标注模型，以及词汇因素与语法因素的形式化方法进行描述。

3.1 隐马尔科夫模型（HMM）

HMM 可以表述为：

$$\hat{Q} = \arg\max_{Q} P(Q|W) = \arg\max_{Q} P(Q)P(W|Q)$$

根据马尔科夫的有限历史假设，当前状态只与有限历史的状态序列相关，如果将历史长度设定为 2，则有：

$$P(Q) \approx P(q_0)P(q_1|q_0)\prod_{1<i<h} P(q_i \mid q_{i-1}q_{i-2}) = \frac{count(q_0q_1)}{length(corpus)}\prod_{1<i<h}\frac{count(q_{i-2}q_{i-1}q_i)}{count(q_{i-2}q_{i-1})}$$

其中 length(corpus)是语料库中的词例总数，count(x)是序列 x 在语料库中出现的频数。

从上式可以看出，P(Q) 的算式反映出历史词性序列对于当前词的词性判断作用。由于转移概率只涉及到词性序列之间的约束关系，而不直接涉及词汇信息，因此可以近似地表示句法因素对词性标注的影响程度。

P(W|Q) 是当前可能的词性序列到当前词序列的发射概率，该值可以通过训练语料求得。为了简化 P(W|Q) 的求解过程，我们假设训练语料中的词具有独立性，因此 P(W|Q) 的计算方法可以表示为：

$$P(W|Q) \approx \prod_{i,j} P(w_i|q_j) = \prod_{i,j}\frac{count(w_i^{q_j})}{count(q_j)} \quad (w_i \in W, q_j \in Q)$$

其中 w_i 表示当前待标注的词；q_j 表示 w_i 的一个可能词性；$count(w_i^{q_j})$ 表示在训练语料中词性为 q_j 的 w_i 的频数；$count(q_j)$ 表示训练语料中词性 q_j 的频数。

从上式可见，发射概率是词性和词形的关系，能够较好地反映出待标注词的词汇因素对于词性判断的作用，因此，P(W|Q) 的算式近似地表示了词汇因素对词性标注的影响程度。

由上述 HMM 的算式可知，这一模型标注词性既考虑了语法因素，也考虑了词汇因素。

3.2 马尔科夫模型（MM）

MM 只是利用历史词性序列来推测当前词的可能词性， MM 的基本公式可以表达为：

$$\hat{Q} = \arg\max_{Q} P(Q)$$

P(Q)的计算公式与 HMM 模型相同。

由于 MM 只利用了词性转移概率来标注词性，因此可以认为 MM 只利用了句法因素判定词性。

3.3 词汇最大概率模型（MP）

词汇最大概率模型的基本思想就是为每个待标注的词赋予该词在训练语料中出现概率最高的词性。因此 MP 可以表示为：

$\hat{Q}=q_1\cdots\cdots q_h$，其中 $q_i=\arg\max_{q_j} P(q_j \mid w_i)$ （0<i<h+1）

每个词的各个可能词性在训练语料中出现的概率可以根据训练语料计算得到，具体方法是：

$$P(q_j|w_i)=\frac{count(w_i^{q_j})}{count(w_i)}$$

从公式可以看出，MP 模型与历史词性序列无关，也与语境中的其他词无关，只与当前词的可能词性相关，因此可以认为 MP 模型只利用了词汇因素对词性进行判定。

3.4 各种因素在三种模型中对词性标注的影响

表 1 给出了词的核心语义因素（Sem）、词法因素（Mor）、句法因素（Syn）在三种模型中对词性标注的影响。

表 1 词的核心语义因素、词法因素和句法因素在三种模型中对词性标注的影响

标注模型	Sem	Mor	Syn
HMM	+	+	+
MM	-	-	+
MP	+	+	-

注：“+”表示考虑该因素；“-”表示未考虑该因素。

除了以上列出的语法词汇因素外，非语言因素在三种模型的标注中都对标注准确率产生影响。

用 P(L,M)表示语言 L 用模型 M 进行词性标注的准确率，其中 L 取 E、C 分别表示英语和汉语，M 取 HMM、MM、MP 分别表示隐马模型、马尔科夫模型和词汇最大概率模型；并用 f(L,F)表示语言 L 中因素 F 对词性标注的影响，其中 L 的取值同上，F 取 Sem、Mor、Syn、Nfs 分别表示词的核心语义因素、词法因素、句法因素、非语言因素对于词性标注的影响，根据上表和前面的分析，就大致地有以下关系式：

P(E, HMM)~f(E, Sem)+f(E, Mor)+f(E, Syn)+f(E, Nfs)

P(E, MM)~f(E, Syn)+f(E, Nfs)

P(E, MP)~f(E, Sem)+f(E, Mor)+f(E, Nfs)

P(C, HMM)~f(C, Sem)+f(C , Syn)+f(C , Nfs)

P(C, MM)~f(C, Syn)+f(C, Nfs)

P(C, MP)~f(C, Sem)+f(C, Nfs)

其中“~”表示公式两边的相关性。C 没有词形变化，所以 P(C, HMM)比 P(E, HMM)少一项影响因素 f(E, Mor)，P(C, MP)比 P(E, MP)也少一项影响因素 f(E, Mor)。

从上面的关系式可以看出不同语言和不同模型在词性标注准确率中反映的相关因素的差别。

表 2 模型差异和语言差异在相关因素中的反映

相关因素 / 模型差 / 语言	HMM-MM	HMM-MP
英语	f(E, Sem)+f(E, Mor)	f(E, Syn)
汉语	f(C, Sem)	f(C, Syn)

从表 2 中看出，

P(E, HMM)同 P(E, MM)的差值大致反映了英语词的核心语义和词法因素对词性标注准确率的影响，亦即英语词汇因素对词性标注准确率的影响。

P(E, HMM)同 P(E, MP)的差值大致反映了英语的句法因素对词性标注准确率的影响。

P(C, HMM)同 P(C, MM)的差值大致反映了汉语词的核心语义因素对词性标注准确率的影响。

P(C, HMM)同 P(C, MP)的差值反映了汉语的句法因素对词性标注准确率的影响。

需要注意的是，由于非语言因素的影响存在于各类标注模型之中，本研究采取计算不同模型的标注准确率差值的方法将非语言因素对词性标注的影响大致抵消掉，更好地考察语言因素对词性标注的影响。

词性标注准确率与各种因素的关系是复杂的，难以严格准确地量化。本文求两种模型对同一种语言的标注准确率差值，一种方法是用准确率的绝对差值，即所差的百分点数；另一种方法是用二者的相对差值。模型 M1 和 M2 对语言 L 的词性标注准确率的相对差值为：

$$\frac{P(L,M_1)-P(L,M_2)}{1-P(L,M_2)} \quad (P(L,M_1)>P(L,M_2))$$

相对差值反映了准确率提高的价值。对于同一个绝对差，在已经很高的 $P(L,M_2)$ 基础上提高，比起在较低的基础上提高，价值要大一些。

本节的分析体现为表 2 中的因素加减，这种加减虽然是示意性的，但仍然可以大致反映这些因素的影响程度。

4 实验设计

4.1 语料及预加工

汉语训练语料选自北京大学计算语言所加工的 2000 年 2 月《人民日报》标注语料，测试语料选用 2000 年 1 月《人民日报》的部分标注语料。

表 3 汉语词性标注的训练语料与测试语料划分

语料类别	语料内容	语料规模（tokens）	语料规模（types）
训练语料	2000 年 2 月人民日报	1050934	36661
开放测试语料	部分 2000 年 1 月人民日报	144912	12667

英语训练语料选用宾州树库华尔街日报语料的 00-19 组，测试语料选用 23-24 组。

表 4 WSJ 训练语料与测试语料的划分

语料类别	语料内容	文档数	语料规模（tokens）	语料规模（types）
训练语料	WSJ 00-19 组	2000	1012809	29739
开放测试语料	WSJ 23-24 组	200	108540	8161

4.2 实验和评价方法

我们对英语、汉语、变形英语的上述语料，用 HMM、MM、MP 三种模型分别进行词性标注，计算标注准确率，进而计算 P(L,HMM)-P(L,MM)和 P(L,HMM)-P(L,MP)，用以估计各种因素对语言 L 的词性标注准确率的影响。同时，我们还对兼类词标注准确率进行了统计，并利用以上指标进行对比分析。

5 实验结果

5.1 总体标注准确率的比较

表中 HMM、MM、MP 列前两行的数据分别是英语和汉语使用相应模型进行词性标注的总体准确率，第三行是准确率差的百分点数。HMM-MM、HMM-MP、MP-MM 这三列的前两行是对

应模型准确率的差，斜杠左边是差的百分点数，右边是相对差值，第三行是前两行对应值的差。

表5　总体标注准确率的比较

	HMM	MM	MP	HMM-MM	HMM-MP
英语	96.36%	93.66%	93.84%	2.70/0.43	2.52/0.41
汉语	94.85%	88.33%	92.53%	6.52/0.56	2.32/0.31
英语-汉语	1.51	5.33	1.31	-3.82/-0.07	0.20/0.10

从表5并对照表2可以看出：

（1）从总体标注准确率看，英语使用三种模型的标注准确率都普遍高于汉语标注准确率，这是由多种因素造成的，既包括语法与词汇等语言因素，也包括训练语料与测试语料的规模、质量等非语言因素。

（2）从HMM-MM一栏看，英语与汉语使用MM模型标注准确率都较HMM模型有所降低，这是词汇因素缺失所致，但是汉语的降低幅度显著高于英语的降低幅度，绝对差值达到3.82个百分点。可见词汇因素对汉语词性标注的影响显著大于对英语词性标注的影响。

（3）从HMM-MP一栏看，英语与汉语使用MP模型的标注准确率都较HMM模型有所降低，这是句法因素缺失所致。但是英语的下降幅度大于汉语，英语的绝对差值高出汉语0.2个百分点，尽管绝对差值并不大，但英语的相对下降幅度（0.41）高出汉语（0.31）10个百分点，相对下降幅度明显。

（4）单独考察英语的标注结果，可以发现英语应用MM模型的标注准确率与应用MP模型的标注准确率基本相仿，绝对差值不到0.2个百分点，表明在英语词性标注中，词汇因素与句法因素对词性标注的影响基本相仿。

（5）单独考察汉语的标注结果，可以发现汉语应用MM模型的标注准确率要显著低于MP的准确率，相差4.2个百分点，表明在汉语词性标注中，当只考虑句法因素而排除词的核心语义因素时，其标注准确率要大大低于只考虑词的核心语义因素而排除句法因素时的标注准确率。

5.2　兼类词标注准确率的比较

下表给出了不同模型的兼类词标注准确率及差值比较结果。

表6　兼类词标注准确率的比较

	HMM	MM	MP	HMM-MM	HMM-MP
英语	94.63%	88.97%	90.10%	5.66/0.51	4.53/0.46
汉语	88.12%	68.92%	81.20%	19.20/0.62	6.92/0.37
英语-汉语	6.51	20.05	8.90	-13.54/-0.11	-2.39/0.09

该表的说明可参见表5。

从表6看出：

（1）英语和汉语的词性标注准确率都有较大下降，因为这里排除了非兼类词的影响，指标低是正常的。汉语下降幅度明显超过英语，说明这些模型与汉语的词性标注任务适配性不大好。

（2）表6中基本上所有表示差值的数都与表5符号相同而数值明显扩大，这就强化了上面根据表5分析出的各项结论。唯一符号不同的是HMM-MP栏目，从绝对差值上看英语值低汉语值高，但相对差值还是英语值高于汉语，仍说明句法因素对英语词性标注的作用超过汉语。

5.3　排除词法信息后的总体标注准确率

为了进一步考察词法因素对英语词性标注的影响，我们对英语的训练语料和测试语料都做了

词形还原处理，形成了变形的英语 E*，并利用处理后的语料进行了标注实验。实验结果如下：

表 7　排除词法信息后的总体标注准确率的比较

	HMM	MM	MP
变形英语 E*	90.18%	83.54%	84.28%
英语 E	96.36%	93.66%	93.84%
汉语 C	94.85%	88.33%	92.53%
E-E*（百分点）	6.18	10.12	9.56
E*-C（百分点）	-4.67	-4.79	-8.25

从以上数据可以看出，对英语进行词性还原处理后，每个模型的准确率都大幅度降低，可见英语词法对词性有重要作用。降低后的准确率比汉语低不少，说明英语若没有词法变化，它的词性歧义比汉语更严重。

5.4　词汇因素与语法因素在英汉语词性标注中作用的总结

从三种衡量方法的英汉对比中可以看出：

（1）汉语中的词汇因素对于词性标注的影响要显著大于英语。

（2）英语中的语法因素（包括句法因素与词法因素）对于词性标注的影响要显著大于汉语。

（3）在英汉语内部，英语的词的核心语义因素与句法因素对词性标注的影响基本相仿；而汉语的词的核心语义因素对词性标注的影响显著高于句法因素。

6　讨论与展望

本文的研究主要基于自动词性标注实验，通过比较不同标注模型的标注差异，分析句法因素、词法因素以及词的核心语义因素在英语与汉语词性标注中的作用，从而探究汉语词类及词性标注的特点。从实验结果分析看出，汉语的词的核心语义因素对于汉语词性标注起着关键作用，这与英语十分不同。

本文中说的句法因素是基于简化的句法模型的，其实只是 N-gram 词序因素。虽然词序是句法的重要方面，但毕竟还不是通常研究的带有层次的短语结构的句法。区别词序因素和短语结构句法因素对词性标注的影响，研究英语和汉语的词性与短语结构句法的关系，将是我们的进一步要做的工作。

参 考 文 献

[1] 洪堡特. 洪堡特语言哲学文集[C]. 姚小平 译. 湖南：湖南教育出版社，2001.

[2] Bisang, Walter. Precategoriality and Syntax-based Parts of Speech: The Case of Late Archaic Chinese. Studies in Language[C], 32.3 (2008): 568-589.

[3] 陈小荷. 从自动句法分析角度看汉语词类问题[J]. 语言教学与研究，1999.

[4] 董振东、董强. 面向信息处理的词汇语义研究中的若干问题[J]. 语言文字应用，2001.

[5] 郭锐（2002）《现代汉语词类研究》，商务印书馆，北京

[6] 陆俭明. 现代汉语语法研究教程[M]. 北京：北京大学出版社，2005.

[7] 潘文国. 汉英语对比纲要[M]. 北京：北京语言大学出版社，1997.

[8] 宋柔. 从语言工程看汉语词类[C].《语言学论丛》(第四十辑)，2009.

[9] 宋柔，邢富坤. 再从语言工程看汉语词类[C].《语言学论丛》(第四十四辑)，2011.

[10] 邢富坤. 现代汉语词类体系与词性标注研究. 北京语言大学博士论文，2010.

[11] 俞士汶，朱学锋，王惠，张芸芸（1998）《现代汉语语法信息词典详解》，清华大学出版社，北京

[12] 朱德熙. 语法答问[M]. 北京：商务印书馆，1985.