文章编号： 00-0000-00

上古汉语分词及词性标注语料库的构建
--以《淮南子》为范例*

留金腾[1、2]，宋彦[1]，夏飞[3]
(1. 香港城市大学中文、翻译及语言学系，香港九龙达之路 83 号；
2. 香港理工大学香港专上学院，香港九龙油麻地海庭道 9 号；
3. 华盛顿大学语言学系，美国华盛顿州西雅图，邮箱 354340，邮编 98195)

摘要：本文介绍了以《淮南子》为文本的上古汉语分词及词性标注语料库及其构建过程。我们采取了自动分词与词性标注并结合人工校正的方法构建该语料库，其中自动过程使用领域适应方法优化标注模型，在分词和词性标注上均显著提升了标注性能。我们分析了上古汉语的词汇特点，并以此为基础描述了一些显式的词汇形态特征，将其运用于我们的自动分词及词性标注中，特别对词性标注系统带来了有效帮助。我们总结并分析了自动分词和词性标注中出现的错误，最后描述了整个语料库的词汇和词性分布特点。我们提出的方法在《淮南子》的标注过程中得到了验证，为日后扩展到其它古汉语资源提供了参考。同时，基于本文工作得到的《淮南子》语料库也为日后的古汉语研究提供了有益的资源。

关键词：上古汉语语料库；分词；词性标注；领域适应
中图分类号：TP391　　**文献标识码：**A

The Construction of a Segmented and Part-of-speech Tagged Archaic Chinese Corpus：A Case Study on *Huainanzi*

Kam Tang Lau[1, 2], Yan Song[1], Fei Xia[3]
(1.Department of Chinese, Translation & Linguistics, City University of Hong Kong,
83 Tat Chee Ave., Kowloon, Hong Kong SAR, China;
2. Hong Kong Community College, The Hong Kong Polytechnic University,
9 Hoi Ting Road, Yau Ma Tei, Kowloon, Hong Kong SAR, China
3. Department of Linguistics, University of Washington,
PO Box 354340, Seattle, WA 98195, USA)

Abstract: In this paper, we present a segmented and part-of-speech (POS) tagged Archaic Chinese corpus along with its construction process, which is performed by automatic segmentation and tagging with manual correction as post-processing. We use both Modern and Archaic Chinese labeled data for training word segmenter and POS tagger, which are further improved by domain adaptation techniques, as well as by adding linguistic and morphological features derived from the characteristics of Archaic Chinese language. The experimental results showed the effectiveness of our approach. In particular, the domain adaptation techniques and the added features significantly improve POS tagging performance. During our manual correction, we categorize the errors resulted from the automatic segmentation and POS tagging process, and investigate the sources of those errors. Finally, we give the statistics of the resulted corpus on the distributions of words and POS tags. Our work is a preliminary study that could be easily extended to annotating other Archaic Chinese text, and the resulted corpus is a valuable resource for research on archaic Chinese language.

Key words: Archaic Chinese Corpus; Word Segmentation; Part-of-speech Tagging; Domain Adaptation

---

* **收稿日期：**　　**定稿日期：**
基金项目：基金名称（基金编号）；基金名称（基金编号）
作者简介：留金腾（1980—），男，博士研究生，助理讲师，主要研究方向为训诂学及词汇学；宋彦（1981—），男，博士研究生，主要研究方向为计算语言学；夏飞（1971—），女，副教授，研究方向为计算语言学、自然语言处理及语料库建设。

1. 引言

在互联网以及数字化浪潮的推动下，大量的文本资源变得易于获取，从而推动了语料库技术的发展。为了在词汇学，语义学和语法学等方面深入研究，人们构建了大量带标注语料库（Labeled Corpus）以提供丰富的描述信息以供进行多层次的分析和检索。然而，由于实际应用的需要，这些语料库大多都是基于现代语言；以汉语为例，多数语料库都来自近二十年的新闻及通讯文本（例如，[12][17]）。在古汉语[1]方面，这类带标注的语料库资源极度匮乏。随着汉语研究的深入，对于这类资源的需求也变得更为迫切。由于古汉语相对于现代汉语有很多不同之处，无法简单地把现代汉语的资源运用于古汉语的研究和分析，建立专门的古汉语语料库尤为重要。

在此基础上，我们选择以上古汉语文献《淮南子》为基础，建构了一个上古汉语分词及词性标注[2]语料库。我们分析了上古汉语的一些特点，包括构词，形态和语素等方面，并针对这些特点提出了可能影响分词和词性标注的问题。进而，为了有效提升语料库构建的效率，并且尽量降低人工标注的工作量和主观标注错误率，我们构建语料库的过程采用了自动标注与人工校正相结合的方法，首先利用自动方法为语料切分词语并进行词性标注，然后在自动标注的基础上进行人工校正。值得提出的是，我们的自动分词和标注方法有效利用了现代汉语的带标注资源，并使用领域适应（Domain Adaptation）方法显著提高了分词和标注的准确率，降低了后续的人工校正工作量。在人工校正的同时，我们还分析了自动标注存在的问题，并总结了不同的错误类型，为改进后续的自动方法提供了指导。最终，我们仅使用较少的人力和时间便得到了一个拥有接近 14 万词规模的高质量上古汉语分词及词性标注语料库，这对于古代汉语词汇研究及解读经典，有莫大的助益。同时，以标注得到的语料库为基础，可以更方便地对更多古汉语资源进行分析和处理。

本文的结构描述如下：第二节为相关工作，我们选取并描述了两个具有代表性的语料；第三节是以《淮南子》为基础，针对上古汉语的语言特点的分析，并从中提出了一些有助于自动分词和标注的特征；第四节描述了我们的方法，特别是采用领域适应技术对上古汉语进行分词和标注的量化分析，同时总结了人工校正过程中发现的分词及标注问题；第五节是该语料库的词汇及词性统计分析；最后一节是对本文的总结。

2. 相关古汉语语料库简介

目前，上古汉语语料库资源还较为稀少，特别是带标注的语料库尤其稀缺。以下我们介绍两个（分别为带标注和不带标注）现存较大规模、覆盖较为全面的古代汉语语料库。

2.1 台湾中央研究院上古汉语标记语料库（Academia Sinica Tagged Corpus of Old Chinese）

该语料库是台湾中央研究院古汉语语料库（Academia Sinica Ancient Chinese Corpus）的次级语料库[1]，从 1995 年开始输入及标注，计划输入包括先秦至西汉时期七十多种的文献，现今完成并开放使用的有三十六种文献，其中包括十三经和先秦诸子，以及一部分西汉的著作。该上古语料库包括大约 2,500,000 汉字，是现今最大的上古汉语带标记语料库。

该语料库进行了分词和词性标注，并以检索为主要使用目的，提供了以词为单位的线上检索功能。因此，检索结果对于分词的准确性和一致性较为敏感。如果出现切分甚至标注不一致的情况，则可能无法得到检索结果[3]。另一方面，该语料库采用的是其自定义的词性标注规范，与现代汉语的流行标注规范[12]有较大差别，如果需要对古今汉语进行比较研究，会带来一定困难。

---

[1]根据文献[10]的说法，汉语词汇史分期，可以东汉为界，在大约公元 3 世纪以前的是上古汉语；东汉其下的是中古汉语；南宋（大约公元 13 世纪）之后，则是近代汉语；五四运动（公元 1919 年）以来，就是现代汉语。

[2]分词及词性标注是最基本的语料库标注信息。实际上，有了分词和词性标注信息，很多古汉语的相关研究都可以得到很大的帮助。

[3]例如，我们检索“司马”及“司马子反”，分别得到：“酣战之时，司马(NA1)[+others]子反(NB1)[+prop]渴而求饮””以及“临战，司马子反(NB1)[+prop]渴而求饮”其中关于“司马子反”的两个记录具有不同的标注。“司马”是战国时期的官职，并不是罕见词，该语料库针对“司马”一词具有不一致的切分结果，其中前一个切分和标注更为合理。

2.2 北京大学中国语言学研究中心古代汉语语料库（Center for Chinese Linguistics PKU Corpus）

该语料库[2]收录了先秦至民国的古代汉语语料，包括经、史、子、集各类文献，主要有《十三经》、二十五史、诸子百家、《全唐诗》、《全宋词》、《全元曲》、《道藏》、《大藏经》等，超过 1 亿 7 千万字，可能是现存最大可供在线检索的古代汉语语料库。

尽管该语料库规模较大，但该语料库并未进行分词及词性标注，所以只能提供给学者检索文献及语句之用，在检索时可以设定最多显示字数，并且可自定检索字词左、右的字数，提供窗口模式的检索结果。例如检索“羽民”，并设定左、右各显示 10 个字的窗口，就会出现如下的结果：

表 1 北大中国语言学研究中心古代汉语语料库检索结果

有角，乘之寿二千岁。	羽民	国，其民皆生羽毛。卵
比翼。一曰在南山东。	羽民	国在其东南，其为人长
奔晋。 又【山海经】	羽民	国，其人长项，身生羽
仆，兽名。【博物志】	羽民	国有兽，文似豹，名虎

该语料库为古代汉语研究提供了很大的帮助。然而，因为没有分词和词性标注信息，对古代汉语词汇和语法的研究帮助有限，如若对该语料进行分词和词性标注，进一步提升语料库的应用价值，对古汉语文学、语言、文化研究将更有裨益。

3. 《淮南子》及上古汉语词语的特点

我们的工作使用《淮南子》一书作为标注语料。《淮南子》是西汉淮南王刘安（前 179－前 122）及其门客集体撰写的一部著作，是模仿《吕氏春秋》而编撰的，其内容广博，包览各种知识，其词汇丰富，并且具有多样性，是上古汉语比较有代表性的作品[4]。因此，以本书作为语料可以保证我们得到的带标注语料库对上古汉语具有较高的覆盖度，并且更方便将来以此为基础，将标注资源扩展到不同的上古汉语文献。这一节我们将介绍一些基于我们的预料中上古汉语中可能对分词和词性标注产生影响的特点。

3.1. 古汉语复音词构词特点

通常情况下，在上古汉语中，一个字便可以独立成词。然而，除了这些单音词之外，上古汉语也已经有了复音词[10]，这些词在构成后，每个单字变成词的一部分，这些字构成的复音词很容易产生词语切分的歧义。同时，这些构成复音词的字，即使本来有各自的意义，一旦成为复音词后，它们本身的词义便可能与新形成的词义不同，因此一定程度上也会带来词性标注的歧义。例如“春”和“秋”本来是两个季节，如《管子・形势解》：“故曰：春/秋/冬/夏/，/不/更/其/节/也。”合成一词则应解释为“年”，也代表年岁，如《战国策・秦策》“王/之/春秋/高/，/一/日/山陵/崩。”而后又变为时代的称呼：周代的春秋时期，如《论衡・龙虚》“春秋/之/时/，/龙/见/于/绛/郊。”这类单音复音同步使用的现象在上古汉语中非常普遍，然而在现代汉语中却一般较为少见，前述单音词分开使用时，一般要加上词缀，如“天”、“季”等。下面我们列出上古汉语中的复音词的几种形式，并指出他们可能带来的分词或者词性标注的困难。

3.1.1 构成新义

由本身有意义的语素组成，形成新的概念，表达一个与原本意义不同的词。如：左右（君主的近臣）、执圭（战国时代楚国的一个爵位）、股肱（君主的得力大臣）。这一类词在标注的时候，需要特别注意意义的变化所带来的词性转变。

---

[4]西汉（前 206－9）建立了中国历史上繁荣的统一国家，其科技、文化相比于先秦时期有了较大的发展，社会生活更为丰盛，促使语言词汇的丰富与发展，新的词语不断产生，形成新的语言面貌。而且，西汉是上古汉语向中古汉语发展的阶段，学者一般认为汉语复音化是在东汉时开始迅速加快[3][5]，而《淮南子》作为西汉文献，介乎先秦汉语与东汉汉语之间，可以预见，这一时期的汉语既逐渐增加复音词汇，又保留较多先秦至汉初的上古汉语词汇，其词汇内容是很丰富的。

3.1.2 并列复合

由意义相同、相近，甚至相反的语素组成，它们在单独使用和合成使用的意义都没有变化，一般合成使用时，是为了配合句子的音节，也可以说是为了配合语法的需要而结合的，如：布施、积聚、寂漠、仁义、礼乐。而反义复合词有的是同时总括两面的意义，并非实指，如赏罚、进退。

3.1.3 偏正复合

同现代汉语的偏正结构，前一个语素修饰后一个语素，形成的意义一般没有变化，通常也可以将其中各个单音字分开使用，然而由于其结构和意义很固定，使用频率又比较高，所以也当视作复音词。如：明主、道术、美人等。

3.1.4 意义偏指

由同类或反义的语素组成，结合后只保留其中一部分的意义，或者是淡化另一部分的意义。如：禽兽（只表示兽类）、国家（只表示诸侯封地“国”，不表示大夫封地“家”）、肌肤（表示皮肤，没有肌肉之义）、无有（指“无”）、好恶（指“好”）。

3.1.5 特指和泛指的转化

由本有所特指的语素组成，本身意义没有变化，但整体的意义指涉更宽泛了。如：骨骸（本指骨头，转化为泛指身体）、江河（本指长江和黄河，后则泛指河流，由此会产生词性从专有名词到一般名词的变化）。另外还有由本来泛指的意义，合并后转为特定的意义。如：北面（称臣的特指，从 LC 到 VV 的词性变化）、三王（特指远古三皇[5]，一般指伏羲、神农、女娲）、五帝（特指上古五帝，一般指黄帝、颛顼、帝喾、帝尧、帝舜）。

另外，上古汉语还存在一些单纯复音词，主要有：1．叠音词，如：昭昭、旷旷；2. 双声词，如：蟋蟀、蔟蒙；3. 叠韵词，如：笼蒙、常羊等。除此之外，一般的古代漢語都應視為單音詞，即使有些使用頻高，意義緊密的兩字詞，我們也只能視為是詞組，而不是詞語，如“異類”、“无端”等。而有的是在古漢語的語法下，不能成詞，如“喜不以賞賜”，是说用“賞”（礼物）来“賜”（赠送）给别人，而并非把“赏赐”当作为一个词语。正因为大多数单音词以及上述一部分复音词的存在，给词语切分带来了一定困难。

3.2 上古汉语词语形态特征

除了复音词的构成特点，上古汉语在词语的组成模式方面，也存在一些形态特征。基于上一节所描述的复音词的构词特点，我们更倾向于寻找一些显式的词语形态特征，这些特征将更容易被归纳并使用到自动分词和词性标注中。

3.2.1 词语形态

在上古汉语中，有一部份复音词是由同义、近义或者同类形的单音字（或单音词）组成的。这一特点反映在上古汉语中，汉字中同一偏旁的字，一般具有近义或同义关系。我们找出《淮南子》书中一些相同偏旁并组成复合词的例子，如表 2 所示。我们希望这种特征可以帮助分析一些复音词语的构成。

表 2 同偏旁词语示例

偏旁	词例
𧾷 /足	跂跃、踊跃、踌躇、蹐蹈、跳跃
木	桃李、梁柱、梧槚、棺椁、杨桃
忄/心	憚悇、憔悴、忼慨、怳忽

同时，利用偏旁的性质，也可以帮助自动标注过程更准确地判定词性。依照偏旁的性质，可以一定程度上判定是词性的大类，例如动词或者名词。以上述例子来看，“𧾷 /足”部的复合词，一般都是动词，而具有“木”偏旁的，都是名词，甚至可以分出凡“木”部的都是树名、木材或木制器具。“忄/心”部的则大多表示心情的状态，较多是形容词。实

[5]上古汉语中，“三皇”常常写作“三王”。如《礼记・内则》：“凡养老，五帝宪，三王有乞言。”《战国策・秦策一》：“虽古五帝、三王、五伯，明主贤君，常欲坐而致之，其势不能，故以战续之。”《庄子・天运》：“夫三王五帝之治天下不同，其系声名一也。”。

际上，这一特征在现代汉语中也适用[11]，在现代汉语的词性标注实验中也证实了该特征的有效性。

3.2.2 词语模式

在上古汉语中，一种非常普遍的词语模式是叠字组合[6]。这些叠字组合很多情况下都应当被当作一个词。此外两组叠字一般也是词，如果结合上述的偏旁来看，同偏旁的 AABB 叠字，则皆是词。例如：睢睢盱盱、昧昧瞵瞵、浩浩荡荡、洞洞瀻瀻等词。而且，在我们的语料中，这类词语通常具有形容词属性，因此这类词语模式特征，可以对自动和人工词性标注带来帮助。

3.2.3 前后缀特征

上古汉语中有些词缀相对固定，并和它们之前或之后的字组成词语。通常某个单字和这些词缀组成的词语代表着某一类事物的意思，例如“者”[7]，和它组成词语以表示某种人，如“狂者”，也可以代表某种物类，如“羽者”表示鸟类，“毛者”表示兽类。

表 3 古汉语词缀示例

词缀	意义	词例
者	人类、物类	射者、罗者、渔者、羽者、毛者
然	情态、样貌	漠然、澹然、恬然、肃然、寂然
氏	人名	伏牺氏、夏后氏、容成氏、中行氏、匠骊氏
公	有爵位之人	晋献公、齐桓公、齐庄公、秦穆公、晋平公
伯	有爵位之人	智伯、穆伯、郑伯、郈昭伯、中行缪伯

这些词缀对词性标注较为有益，例如凡缀有“者”、“氏”、“公”、“伯”等字的复音词，大多可以标为名词。其外，缀有“然”字的复音词，如果其后是动词的话，这个“V＋然”构成的词通常便是副词。

3.3 古汉语的词性转化

古汉语的词性常常因应他们的语法位置而改变[10]，而常见的词性转化情况有两种：兼类和活用[8][8]，我们在做词性标注时，要因应相关的具体情况作出判断，才能保证词性标注的准确性。

3.3.1 兼类

兼类是说一个词经常具有两个不同词类（性）的功能[8]，这种情况在古汉语中大量存在，一般来说，这种兼用的汉语词汇，大部分体现在名词、动词兼用上，也延续到现代汉语里，例如“效”，现代汉语因其动词和名词词性而有“效劳”和“功效”二词。下面举《淮南子》中的例子，其中 NN，VV 和 JJ 分别为名词、动词和形容词。[9]

表 4 古汉语兼类词语示例

词语	词性	例句	意义
朝	NN	乃赏军率武人于朝。	朝廷
	VV	朝成汤之庙，解箕子之囚。	朝拜
言	NN	不能通其言，教俗殊也。	语言
	VV	非道不言，非义不行。	说
幼	JJ	武王崩，成王幼少。	年纪小
	NN	百姓攜幼扶老。	年幼的人

[6]实际上，现代汉语也满足该特征。

[7]这里补充一下，“者”字有时并不作词缀用，而是和“也”字一起便用，用于判定句。例如《淮南子·缪称训》：“道者，物之所导也；德者，性之所扶也；仁者，积恩之见证也；义者，比于人心而合于众适者也。”此处的“……者……也”是一组判定句，当中的“者”字并不能视为词缀。

[8]在文献[8]中，李佐丰还提出一种转称，他认为转称跟兼类不同，主要表现在词义上，兼类无论是动词还是名词，其意义都是明确的，然而转称时所表示的词义并不确定，经常具有转称作用的是形容词。在使用时仍然是形容词，只是从陈述变为指称。在本文中，我们认为没有必要再分出一类转称，因为如果使用较为固定的，可以归为兼用，如果是比较临时的，就是活用。

[9]这些词性均使用自文献[14]所描述的词性标注体系。

兼类词语的标注，需要依据他们的位置和语法功能做出判断。对于这类词语的词性标注，有助于在将来对其所处文本进行句法分析。

3.3.2 活用

活用不是某种词类的固有用法，只是其偶尔出现的一种特例[8]，这种临时改变词性的做法，在古汉语中非常常见。例如：

表 5 活用词语举例

词语	词性	例句	意义
新	JJ	邯郸师有出新曲者。	新的
	VV（活用）	弊而复新，其为乐也。	使之变新
西	LC	正西弇州曰并土。	西方
	VV（活用）	将欲西而示之以东。	往西方去
镜	NN	夫照镜见眸子，微察秋毫。	镜子
	VV（活用）	抱大圣之心，以镜万物之情。	洞察、观照

表中“新”本为形容词（JJ），“西”是方位词（LC），“镜”乃一般名词（NN），但活用时可以都可以做一般动词（VV）。因此，词的活用对于自动标注而言较为困难，通常需要上下文的词性序列帮助判定其活用词性。而对于人工标注而言，可以通过观察其句法位置帮助判定词性。

4. 语料库的标注及校正

我们采用宾州中文树库（CTB）[12]作为我们语料库的分词和词性标注标准[13][14][10]。我们的工作分为两部分，自动标注以及人工校正。在词语切分和词性标注两个环节，自动标注和人工校正是交替进行的，其流程可以简单描述为：

下面分别就自动分词和标注的过程以及人工校正中发现的错误及总结分别描述我们的工作。

4.1 自动工作

针对自动分词和词性标注，我们面临几大挑战：1，理想情况下，我们应该使用已有的古汉语标注语料作为训练数据，而这并不容易得到，甚至因为标注标准不同而无法使用；2，由于没有现成的上古汉语训练语料，如果使用不太相关的训练数据，在古汉语上得到的结果很可能并不理想，进而加大后续的人工标注的难度。因此，我们需要在数据和方法上有效适应这一特殊的应用。考虑到上述问题，我们尝试使用领域适应（Domain Adaptation）方法，并以一定量来自目标领域的种子数据（Seed Data）为基础，借助现代汉语资源，有效地提高分词和词性标注的准确率。

因此，在训练语料方面，我们采用整个宾州中文树库 7.0 版（CTB7）作为训练基线（Baseline）模型的资源。针对前面提到的挑战，我们考虑到：1，由于我们没有其他现成的古汉语标注资源作为训练数据，因此使用已知较好的现代汉语资源是有效且易于推广的方法，同时配合领域适应技术，可以一定程度上解决语料差异带来的负面影响；2，由于我们使用 CTB7 的切分和标注标准，采用 CTB7 作为标注数据不会造成标注结果偏向于其他训练语料的标注标准，从而使得后续的校正和分析亦会更容易；而且，就我们所知，目前也没有使用 CTB7 标注标准的古汉语资料可供使用。对于最终的分词和词性标注，我们使用的种子数据是从古汉语语料中随机选取并人工标注的非常有限规模的一个数据集，用于提供基本的字词特征作为训练数据。

[10]虽然这些规范是针对现代汉语提出的，但对于古汉语的处理依然具有指导意义。

为了较为准确地评估自动标注的效果，我们随机选取了占整个数据约 10%的子集作为测试语料。值得说明的是，这 10%的数据是从《淮南子》各章节中选出的，包括了哲学、政治、人文、建筑、天文、地理、动物、植物等各个主题，有效涵盖了整个语料中的大部分词汇，以这部分数据作为测试语料，所得出的结果，将可以更有效地反映出我们的自动标注系统的性能。针对这部分语料我们完全使用人工分词和标注，以保证其质量，并且这样做会有效避免自动标注的结果带来的偏向性。人工标注的种子和测试数据的统计信息如下表所示。

表 6 训练及测试数据的统计信息

数据	字数	词数	句数
训练（种子）	1,486	1,278	50
测试	15,340	13,133	600

在领域适应方面，我们采用[9]所描述的半监督学习（Semi-supervised Learning）领域适应技术将基于现代汉语训练的模型应用于古汉语的分词任务中。其核心是采用描述长度增益（Descriptive Length Gain, DLG）对古汉语生语料进行非监督学习（Unsupervised Learning），从而得到的所有可能成词字串，并将其转化为特征加入训练语料中，使得训练的模型倾向于对该类型的测试语料有更强的标注能力。具体地，DLG 是 Kit 等[6]提出的一种基于文本信息量（熵）的文本字（词）串评价方法，并有效地使用在了词汇获取的任务中[7]。具体的方法可参考[9]，我们在此不再赘述。值得提出的是，我们的 DLG 特征来自于所有可用训练数据（CTB7+Seed），而且我们针对[4]提出的基于特征扩增（Feature Augmentation）的领域适应方法，将来自 CTB7 的特征扩增为具有通用领域（General Domain）和源领域（Source Domain）标识的两部分，同时将来自种子数据的特征扩增为具有通用领域和目标领域（Target Domain）标识的两部分，然后在此基础上进行模型的训练，使得模型可以有效地针对不同领域的特征估计不同的权重。相应地，测试数据包含的特征也采用与种子数据一致的扩增方案。

分词方面，我们采用了基于字标注的条件随机场（Conditional Random Fields，CRF）模型，使用了被广泛采用的 6-Tag 标注集（B, B2, B3, M, E, S）[9][15][16]，以及 DLG 特征和上文所描述的古汉语语言学特征。最终的特征模版（Template）如表 7 所示。其中所有的字特征属于基本特征（构成基线系统的特征），其余为附加特征。模式和词缀都属于布尔（Boolean）类型（即满足或者不满足某类标记）的特征，仅仅描述该字（及其上下文）是否可以作为某种类型或者某种词缀。例如，字串“睢睢盱盱”可以匹配上 AABB 模式，则对于赋予每个字相应的特征，如“睢 A /睢 A /盱 B /盱 B”，其余字串无法匹配上的，则标为 Null。对于词缀而言，如字符串“伏牺氏”，我们在“氏”字这一常用词缀上加入词缀特征标记，该字符串对应的词缀特征则为“伏 Null/牺 Null 氏/X”。对于偏旁特征而言，直接标记该字的主要偏旁是何种偏旁即可，例如前文表 2 所示。其中，偏旁和模式特征均来自于当前字本身，而词缀特征则来自于人工总结出的一些词缀例子，如前缀“有”，后缀“氏”、“者”、“公”、“伯”、“然”，等等，如表 3 所示。

表 7 分词模型使用的特征模版

特征描述	特征
一元字特征	C_{-1}，C_0，C_{+1}
二元字特征	$C_{-1}C_0$，C_0C_{+1}，$C_{-1}C_{+1}$
DLG 特征	D_0^1，D_0^2，D_0^3，D_0^4，D_0^5
偏旁特征	R_0
模式特征	P_0
词缀特征	X_0

词性标注方面，我们也采用了 CRF 模型以及与分词类似的特征模版，不同之处在于，使用词而不是字作为基本的标注单元。特征模版如表 8 所示，采用了与分词系统类似的基本特

征和附加特征。这里，模式特征是当前词的模式（例如 AABB 等），词缀特征则是该词的前后缀字（例如，“伏牺氏”的前后缀特征为“伏”和“氏”），而偏旁特征则是前后缀字的偏旁信息。

表 8 词性标注模型使用的特征模版

特征描述	特征
一元词特征	W_{-1}，W_0，W_{+1}
二元词特征	$W_{-1}W_0$，W_0W_{+1}，$W_{-1}W_{+1}$
偏旁特征	R_0
模式特征	P_0
词缀特征	X_0

需要说明的是，我们采用串行的分词＋标注的方案，而不是分词和词性标注的联合解码（Joint Decoding），是基于下面的考虑：1，虽然使用联合解码可以得到更好的分词性能，但在词性标注上的结果却可能并不理想，同时采用这种方法训练模型需要耗费巨大的计算资源（相当于使用了非常多的类别，训练速度很慢），往往周期太长；2，由于采用人工验证的标注方式，分词的结果可以得到迅速验证并更正，而且古汉语词粒度和句长通常较小，人工验证也较为方便。因此在验证的分词基础上再进行词性标注可以得到更好的标注结果。

我们的分词和词性标注实验结果如表 9 和 10 所示，其中，分词结果包含了准确率（Precision），召回率（Recall）和 F 值（F-score），词性标注结果使用标注精确率（Accuracy）来描述。表中基本特征来源指字词特征来自于哪个语料（或者两个语料的并集），DLG，RPX 是指我们的附加特征，Domain 指采用特征扩增方案。为了进一步展现不同系统的差距，我们还使用了错误率降低（Error Rate Reduction, ERR）指数来描述各个系统相对于基线系统[11]的性能提升。

表 9 基于不同训练数据和方法的分词结果[12]

基本特征来源	附加特征类型	F-score	Precision	Recall	ERR on F
CTB7	-	61.24	70.13	54.35	0
CTB7	+DLG	61.33	71.49	53.70	0.23%
CTB7+Seed	-	65.40	73.40	58.98	10.73%
CTB7+Seed	+DLG	69.88	76.81	64.10	22.29%
Seed	-	75.70	71.31	80.67	37.31%
Seed	+DLG	77.14	75.39	78.98	41.02%
Seed	+DLG+RPX	79.96	79.85	80.08	48.30%
Seed	+DLG+RPX+Domain	**83.70**	**80.23**	**87.49**	**57.95%**

表 10 基于不同训练数据和方法的词性标注结果[13]

基本特征来源	附加特征类型	Accuracy	ERR
CTB7	-	70.44	0
CTB7	+RPX	72.49	6.94%
Seed	-	72.09	5.58%
Seed	+RPX	75.98	18.74%
CTB7+Seed	-	72.87	8.22%
CTB7+Seed	+RPX	76.69	21.14%
CTB7+Seed	+RPX+Domain	**80.81**	**35.08%**

[11]基线系统仅采用 CTB7 作为基本特征的来源，同时不使用任何附加特征。表 9 和表 10 中的第一行分别展示了分词和词性标注基线系统的性能。
[12]其中领域适应方法扩增了 DLG 和 X 特征。来自 CTB7 和 Seed 的这些特征分别对应原领域和目标领域。
[13]其中领域适应方法扩增了 RPX 特征，扩增方式同上。

很明显地，表 9 显示，在分词方面基于有限的种子数据训练（指基本特征的来源）得到的模型比基于 CTB7 得到的模型具有更好的性能，这充分反映了古汉语在构词，词性方面与现代汉语的不同。相对于测试数据，CTB7 可以认为是领域外（Out-of-domain）数据，因此即使其规模远大于种子数据，也无法得到更好的性能。其次，采用了半监督学习方法（＋DLG）的领域适应技术可以进一步增强基于任何训练数据训练的分词模型，对其准确率均会带来提升[14]。加入古汉语的语言学和形态特征（RPX）也可以有效提高整体的分词性能。这些特征可以帮助分词系统得到较高的分词准确率，但同时牺牲了召回率。表 10 显示，在词性标注方面[15]，这些特征同样有效提高了系统性能，实际上证实了我们之前对于古汉语词汇的分析，这些词所包含的语言学和形态特征的确表现出非常强的对词性的指导作用[16]。加入这些特征在词性标注方面对系统的输出结果带来了较大的改善，进而减少了后续人工校正的工作量和难度[17]。最后，使用特征扩增的方法之后，不论对于分词还是词性标注系统的性能都带来了可观的增长。系统在考虑了整体训练集的同时，对于关联到不同领域的特征进行了有效区分，在联合使用 CTB7 和种子数据集进行训练的基础上得到了更好的分词和词性标注结果。表 9 和表 10 中所展示的相对错误率的降幅充分说明了我们方法的有效性，同时也更直观地展现出使用不同特征和方法的差别。

4.2 错误分析及人工校正

通过上面的工作，我们得到了基本的分词和词性标注结果，并进行人工校正。在校正过程中，我们总结出了一些常见的自动切分及词性标注错误，分析如下。

4.2.1 词缀的粘合不准确

在自动切分结果中，存在相当一部分词缀和词语的组合切分不准确的情况。以“氏”字为例，在自动分词时只要前面搭配单一的姓氏，一般都能准确地分为一词，例如“陈氏”、“刘氏”、“赵氏”等。但如果前面搭配的是复姓或其它名称，便不能准确地切分。例如“匠骊氏”便会被切分为“匠骊/氏”，“夏后氏”被切分为“夏后/氏”，“仲孙氏”被切分为“仲/孙氏”等。其他词缀也存在类似复音词干扰的情况，其中“然”字最为突出，搭配单字时比较准确，在搭配两字或以上以表示状态时，错误比例较高。

4.2.2 多音词切分不准确

这种错误大多出现在人名或地名的切分中，例如“厘负羁”（人名）会被切分为“厘/负羁”或“厘负/羁”，或和后面的词混在一起切分，分为“厘/负/羁遗”、“厘/负/羁遗”、“厘/负/羁止”等。这种人名或地名的切分错误，可以使用迭代（Iterative）校正和训练的方法，校正其中具有代表性和高覆盖度的部分例子，然后再进行训练和标注，以提高整体正确率。

4.2.3 一般词性标注错误

在词性自动标注结果中，一般的标注错误主要来自于现代汉语训练语料中某种词类的词性所带来的影响。其一是句末词（SP）标注错误，上古汉语的句末词主要有“也”、“乎”、“焉”、“哉”、“矣”等，其中以“也”的标注错误最明显，大多会被标为副词（AD），这是由于在我们所使用的现代汉语训练语料中，“也”通常是作为副词而存在的。另外，“乎”字作为表示疑问或感叹的句末词，则被错误标注为一般动词（VV）[18]；“焉”字则常被错标为一般名词（NN），另外小部分则被错标为一般动词（VV）。这些错

[14]实际上，考虑到古汉语单子词占绝大多数的情况，如果将古汉语统一切分成一个个单字词，甚至可以得到更高的 F 值。然而，我们的目标是语料库建设，单字切分会为正确标注和修正复音词带来困难，而我们也希望使用已有资源对古汉语标注进行有效指导。

[15]采用 CTB7 和种子数据作为联合训练数据（基本特征来源）可以得到比单独使用各自训练数据更好的性能，这一点与分词结果稍有差异，其原因可能来自两个方面：第一，词性标注对于现代汉语和古汉语的差别没有分词那么敏感；第二，我们的种子数据规模有限，不足以反映出其相对于 CTB7 的优势。

[16]限于篇幅，我们在此并未分别测试各个不同的特征对系统性能的影响。

[17]本语料库的人工分词和标注是由研究上古汉语词汇和词义的学者担任。在没有自动标注协助的情况下，对《淮南子》进行分词和词性标注，一个人大约需要九个月时间。在本文提出的方法和实验环境下，校正工作则只需要大约两个月时间，效率得到了大幅提升。

[18]“乎”也有在句子中担当介词（P），但也大多错标为一般动词（VV）。

误皆缘于古今汉语语法和词义差异，错把一些词依现代汉语词性进行标注，还有如“非”，依照 CTB7 标为系动词（VC），然而在古汉语中，“非”除了作系动词外，还有与“不”有相同的用法，应标为副词（AD）。还有如“无”字，通常标为 VE，然而“无”除了表示没有之外，也有与“不”同样做副词的功能。以上有关“非”和“无”的标注错误也可通过校正并重新训练得到更好的自动标注结果。

4.2.5 多义词导致标注错误

在我们自动标注的上古汉语中，有两个特别容易标注错误的词。其一是“之”，该词既是代词（PN），表示他、它、他们等等，又可表示属格关联词“的”（DEG）。例如“秋毫之末”的“之”应该是 DEG，却误标为作为补语（Complementizer）或名词化尾缀（Nominalizer）的 DEC。其二是“为”，这个词一般可以当作系动词（VC）使用，但又有“做”的意思，应标为一般动词（VV）。例如，“与高辛争为帝”中的“为”是 VV，结果被误标为 VC。有两方面的原因导致了这些词语的标注错误，其一是从如前面所述的现代汉语训练语料带来的差异；其二是在古汉语中，如果上下文信息及其标注序列不能对该词的词性进行有效指导，也会带来标注错误。

4.2.6 古今语法差异导致标注错误

这个主要表现在发语词（Literary Auxiliary Particle）上，发语词是句首的语气助词，起引起下文的作用，没有实际意义。这种情况通常不会出现在现代汉语中，所以在现代汉语的训练语料中缺失了这类训练样本，因而在自动词性标注时大多未能准确标注。本语料中所见的发语词有“夫”、“今夫”、“若夫”等，我们暂且将其词性统一标为其他质词（MSP）。

5. 语料库的统计数据

基于上面的工作，我们得到了一个约 16 万字，接近 14 万词的《淮南子》分词及词性标注语料库。接下来，我们从《淮南子》的词汇及词性方面，分析整个语料库的词频、词长和词性标注的统计分布，进一步展示我们得到的语料库。

5.1 高频词分布

我们构建的《淮南子》语料库包含 11,031 词形（Word Type）。其中，除了一些语法功能词，如“之”、“乎”、“者”、“也”、“而”、“则”等，我们特别统计出现频率超过 300 以上的词语，从中大致看出上古汉语中最为常见的词语分布，如表 11 所示。有意思的是，除了大部分单音词，“天下”是唯一使用频率超过 300 的复音词。

表 11 《淮南子》中词频高于 300 的词语及其词性

词	词频	词性	词	词频	词性
不	3,610	AD	一	472	CD
为	1,459	VC/VV	行	467	NN/VV
有	1,122	VE	非	465	AD/NN/VC
无	1,045	AD/VE	道	419	NN/VV
能	814	NN/VV	天下	380	NN
人	777	NN	至	376	JJ/VV
曰	743	VV	必	371	AD
可	666	VV	见	347	VV
知	635	NN/VV	欲	312	NN/VV
生	542	NN/VV	日	306	NN
得	520	NN/VV	在	302	P/VV

5.2 词语长度分布

根据统计，《淮南子》中绝大多数都是单字词，而复音词又以双音词为最多。词长为三字或四字的，通常都是专名（Proper Names）或叠字词，词长超过四个字的词都是数词。词长及词频的详细统计信息如表 12 所示（语料库中的最长词是 10 字词）。总体上，该语料的平均词长度为 1.15。而对于不同词形，复音词的词形数远多于单字词。从单个词形平均使用率（词数 / 词形）数据也可以看出，复音词相比于单字词使用得并不频繁。

表 12 《淮南子》词语长度频率分布

词长	词数	百分比	词形	使用率
1	117,476	85.67	2,898	40.54
2	18,587	13.55	7,483	2.48
3	725	0.53	399	1.82
4	315	0.23	229	1.38
5	17	0.01	15	1.13
> 5	8	0.01	7	1.14
总数	137,128	100	11,031	12.43

5.3 词性标注分布

整个《淮南子》语料库中仅包含 24 种词性，我们将各个词性按照其词数的频率倒序排列，如表 13 所示。其中，名词和动词占据多数，所有类型的名词（NN、NR、NT）加起来共有 38,109 个，而各类动词(VV、VE、VA、VC)总和则是 32,350 个。与现代汉语相比，我们的语料库所包含的词性种类显然更少，这是因为在古代汉语里，有些句法和词性模式并不存在。例如 DEC(补语或名词化尾缀“的”)、DER（现代汉语“得”）、DEV（现代汉语“地”）、AS（Aspect Particle，动态助词，如“着”、“了”、“过”）等。[19]

表 13 《淮南子》词性标注分布

Tag	词性	词数	比例%	Tag	词性	词数	比例%
NN	一般名词	35,510	25.90	MSP	其他质词	1,442	1.05
VV	一般动词	28,043	20.45	CD	数词	1,355	0.99
PU	标点符号	27,123	19.78	VC	系动词	853	0.62
AD	副词	10,504	7.66	LC	方位词	744	0.54
CC	并列连词	6,160	4.49	JJ	形容词	737	0.54
P	介词	4,711	3.44	NT	时间名词	469	0.34
PN	代词	4,298	3.13	M	量词	640	0.47
SP	句末词	4,088	2.98	DT	限定词	427	0.31
DEG	所有格“之”	4,069	2.97	CS	从属连词	359	0.26
NR	专有名词	2,130	1.55	SB	短被动词	7	0.005
VE	动词“有”	1,779	1.30	LB	长被动词	4	0.003
VA	表语形容词	1,675	1.22	ETC	表多数代词	1	0.001
总数		137,128	100				

6. 总结

本文描述了我们构建的基于《淮南子》的上古汉语分词及词性标注语料库，并着重分析了上古汉语在构词及词性方面的一些特点及其在分词和词性标注方面所带来的困难。在该语料库的构建方面，我们采用了自动分词和标注配合以后续人工校正的方法，利用现代汉语作为基线训练语料，并辅助以非常少量的人工标注上古汉语数据，使用领域适应技术

[19]实际上，古代汉语有些词语也不能完全借用现代汉语的标注。例如，古代汉语的“发语词”，可能应该具有独立的词性，但在本文中，我们暂时归为 MSP，这是为了保证我们的语料库与 CTB 使用同一套标注系统。因此我们语料库中包含的词性标注集可以视为 CTB 标注系统的一个子集，方便将来与现代汉语语料库进行比较。

提升自动标注的准确率，在具有高覆盖度的测试集上证明了我们使用方法的有效性。而在人工校正的过程中，我们总结了自动分词和词性标注中出现较多的错误，分析了错误原因，同时针对部分错误也提出了解决方案。最终，我们得到了一个具有分词兼词性标注的上古汉语语料库。在该工作中，人工工作已经被缩减到了最低限度，相比于从零开始分词和标注，我们已经使得人工工作仅仅局限于校正一些特定的错误类别，并且这些错误很容易通过迭代校正及再次训练得到修正，进一步缩减人工校正的工作量和难度。

通过构建《淮南子》全本分词及词性标注语料，我们的方法被证明可以有效运用于古汉语标注，因此可以进一步推广到其他语料库上，从而利用有限的人力资源得到更多的古汉语标注语料。同时，以《淮南子》为基础，我们已经具备一定量的具有高覆盖度的上古汉语标注资源，以其作为训练数据，可以不再依赖于现代汉语资源，并有效提高未来在古汉语上自动分词及标注的准确率。而且，在此基础上，我们未来还将对这些语料标注到句法甚至是语义角色的层级，为古汉语分析及文本建模提供更为完善的标注资源。目前，本文所提到的语料库还在进一步校正和整理，未来我们将会发布该语料库，进一步完善古汉语资源建设。

参考文献

[1] http://app.sinica.edu.tw/cgi-bin/kiwi/akiwi/akiwi.sh

[2] http://ccl.pku.edu.cn:8080/ccl_corpus/

[3] 程湘清. 《论衡》双音词研究[C]. 程湘清主编. 两汉汉语研究. 济南: 山东教育出版社，1992: 262－340.

[4] Hal Daume III. Frustratingly easy domain adaptation[C]. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2007: 256－263.

[5] 方一新. 东汉语料与词汇史研究刍议[J]. 中国语文, 1996 年, 第 2 期, 140－144.

[6] Chunyu Kit and Yorick Wilks. Unsupervised learning of word boundary with description length gain[C]. In *Proceedings of CoNLL-99*, 1999: 1－6.

[7] Chunyu Kit. Unsupervised lexical learning as inductive inference via compression[C]. In J. W. Minett and W. S.Y. Wang, editors, *Language Acquisition, Change and Emergence.* Hong Kong: City University of Hong Kong Press, 2005: 251－296.

[8] 李佐丰. 古代汉语语法学[M]. 北京: 商务印书馆, 2004.

[9] Yan Song and Fei Xia. Using a goodness measurement for domain adaptation: A case study on Chinese word segmentation[C]. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, 2012:3853-3860.

[10] 王力. 汉语史稿[M]. 北京: 中华书局, 1980.

[11] 王丽杰，车万翔，刘挺. 基于 SVMTool 的中文词性标注[J]. 中文信息学报, 2009, 23(4):16－21.

[12] Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fudong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation[C]. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC-2000)*, Athens, Greece, 2000.

[13] Fei Xia. The Segmentation Guidelines for the Penn Chinese Treebank (3.0), IRCS Report 00-06[R], University of Pennsylvania, Oct, 2000.

[14] Fei Xia. The Part-of-Speech Guidelines for the Penn Chinese Treebank (3.0), IRCS Report 00-07[R], University of Pennsylvania, Oct, 2000.

[15] Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition[C]. In The Sixth SIGHAN Workshop on Chinese Language Processing, Hyderabad, India, 2008: 106－111.

[16] Hai Zhao and Chunyu Kit. Integrating unsupervised and supervised word segmentation: The role of goodness measures[J]. *Information Sciences*, 2011, 181(1):163－183.

[17] Zhou, Q. Annotation scheme for Chinese Treebank [J]. Journal of Chinese Information Processing, 2004, 18(4):1－8.