面向初学者的中文信息处理平台构建及应用

朴敏浚

北京大学 中文系，北京 100871

karmalet@163.com

摘 要：随着大数据的趋势，越来越多的语言研究者开始关注和采用基于大规模语料的研究路线。但是，大部分语言学专业学生的统计知识基础相对薄弱，增加了他们在自然语言处理学习与应用方面的难度。因此，为了降低学习者对计算机与统计知识的陌生感，本文构建了一个用户友好的 ICTCLAS 分词/词性标注界面以及一个简单的基于 N 元语言模型的词串生成器。此平台为学习者提供了一个简易的中文文本处理手段，有利于他们理解统计语言模型的基本运作方式。

关键词：中文分词；N 元模型；词串生成；自然语言处理教学

A Chinese Information Processing Platform for Beginners

PARK MinJun

Department of Chinese Language and Literature, Peking University, Beijing 100871

Abstract: There has been a growing interest for large corpus-based research methodology in the field of linguistics recently. However, this approach uses statistical language processing techniques, which can be difficult to utilize for students who lack knowledge in statistics. In order to address this problem, this study develops a user-friendly part-of-speech (POS) tagging interface of ICTCLAS tagger and a random text generator based on n-gram language model. These tools can provide learners with a simple and easy way to pre-process Chinese texts and help them understand the basics of statistical language model.

Key words: Chinese POS tagging; n-gram; text generation; NLP education

1 引言

近年来，基于大规模数据的信息加工与应用潮流正奔腾向前。大数据、云计算、机器学习的兴起已经超越了计算机科学及信息科学领域，延伸到商业、医疗、媒体等各种领域。例如，京东、亚马逊等网购平台在上千万个顾客的购买历史明细的基础上分析它们的购买模式，给个别顾客提供有针对性的购物推荐服务；与计算机科学、统计学等学科融合的生物信息学（Bioinformatics）在基因识别、基因预测等方面提供了一种低成本高效率的新分析方法。因此，很多学术、商业领域已经充分意识到大数据的前景，正在或将要研发基于大数据的研究方法及应用工具。

语言学也不例外。语言学方面出现了以经验主义、实证主义为导向的研究方向，比如语料库语言学（Corpus Linguistics）、计量语言学（Quantitative Linguistics）等。这些新学

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

科的研究方法就是在大量真实文本的统计与分析的基础上归纳出某些语言现象及其动因的。这种自底向上的大规模定量分析离不开计算机铺助的信息处理过程。可是，在韩国等国外汉语学界中，中文信息处理工具的介绍与普及还停留在初级阶段。因此，为了促进基于大规模语料的语言学研究方法及其在汉语教学上的应用，本文构建及介绍两种中文信息处理工具——GUI 界面的分词/词性标注器和词串生成器。

2 分词与词性标注

“词（word）”在自然语言中占据着重要的地位。从语法的角度看，是组成句子的基本单位，通常定义为“能够独立运用的，有意义的最少语法单位”。从认知上讲，“词”不仅是一个句法单位，而且是组织概念的基本范畴。儿童语言习得过程中，“词”的认识与掌握则视为儿童思维及口语发展的第一阶段。

英语和韩语文本中，空格可以作为单词的分隔符，所以词的界限是比较明确的。而汉语是以字为基本书写单位的，词和词之间没有显性的区分标记。汉语里，“词”在“句”与“字”的中间，是个“隐形”的单位，需要运用一定的知识去发现（俞士汶 2003：121）。由于汉语词的界限是不明确的，这使得汉语分词还有一个难题，即“分词歧义”问题。例如，

南京市长江大桥

提高人民生活水平

大部分手工业品

因此，汉语字符串的切分与“词”的识别问题，即“分词（segmentation）”则为中文信息处理的基础与关键。

词性（或词类, part of speech）是根据词的语法分布来划分的类。众所周知，对汉语这种缺乏形态标记的语言，词类的判定并不容易，兼类现象也特别多。按照词性的不同，词义也随之发生变化。例如：

（1）a. 他是总编辑　　b. 他正在编辑这本书

（2）a. 小心过马路　　b. 我看过这篇小说

同一个词形“编辑”在例（1a）中是名词，充当宾语；在（2b）中则为动词，充当述语。例（2）的“过”也在（a）和（b）中分别具有不同的词性，构成的不同的句法结构（词组）。可见，词性标注的作用不限于词义的区分，它是句法分析（parsing）的先决条件。

综上所述，分词（word segmentation）与词性标注（POS tagging）是中文信息处理的基础性任务，只有在这两种处理的基础上，才能进一步作其他处理，如句法、语义分析等。除了机器翻译、信息抽取等工程应用之外，凡是利用计算机的定量分析都无法回避这一环节——分词与词性标注。对于汉语学习者或研究者而言，已标注好的语料（annotated corpus）是最理想的语料。可是，这些标注语料的规模及来源有限，在很多情况下学习者（或研究者）手中的语料应该是生语料（raw corpus）。假如我们想要分析最新网络流行语，那么应该从互联网得到的语料，就是在生语料的基础上进行分词及词性标注。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.1 汉语词法分析系统 ICTCLAS

汉语词法分析系统 ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) 是基于多层隐马模型(hierarchical hidden Markov model, HHMM)的汉语词法分析系统。该系统的功能有：中文分词、词性标注、未登录词识别。根据开发者张华平博士，它分词正确率高达 97.58%，基于角色标注的未登录词识别能取得高于 90%召回率，其中中国人名的识别召回率接近 98%。

除了 ICTCLAS 之外，还有斯坦福 CoreNLP[①]、结巴[②]等国内外开发的分词系统。但是它们的词性标注集与中国汉语学界广泛使用的两种词性标注集（中科院计算所标注集 ICTPOS、北京大学现代汉语语料库基本加工规范）相差较大。ICTCLAS 就支持这两种词性标注集，所以在跟中国境内的研究成果的兼容性方面它具有优越性。

2.2 ICTCLAS 应用上的难点

ICTCLAS 分词系统在动态链接库（dynamic linking library，dll）基础上运行，驱动系统还需要在其他程序语言（C，Java，Python）中调用该动态链接库（nlpir.dll）。虽然 ICTCLAS 提供实例程序，但是对不懂编程的使用者而言，ICTCLAS 的启动及运用会带来不少困惑。面对这个问题，ICTCLAS 提供一些在线演示网站[③]及单机 GUI 界面[④]。可惜的是，它们是以演示目的为开发的临时界面，在使用中会出现异常或限制，例如限制处理文本大小、不支持文本储存、字符编码不正确、版权到期等问题。

2.3 界面设计及功能

我们在 Python 的 GUI 平台（Tkinter）上设计了界面，基本布局如下：

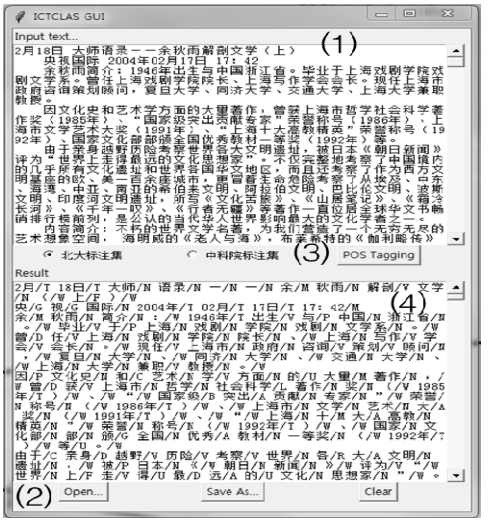

图 1 用户友好的分词/词性标注界面

① http://nlp.stanford.edu/software/segmenter.html
② https://pypi.python.org/pypi/jieba
③ http://ictclas.nlpir.org/nlpir/
④ /bin/ICTCLAS2014/NLPIR_WinDemo.ext

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

使用者可以在输入框（1），直接输入文本。或者通过文件打开的方式（2）读入文本。之后，点击（3）按钮，界面自动分析出文本的编码类型，由 ICTCLAS 分词/词性标注模型完成标注工作，分词结果在（4）框中显示。使用者也可以把这分析结果另存为本地文件。

3 统计语言模型教学的动因及难点

目前自然语言处理的主流方法是基于统计的机器学习。除了上述的自动分词与词性标注之外，句法分析、机器翻译、文本分类等自然语言处理技术或多或少也都建立在统计语言模型的基本理念上。问题是，语言学专业大部分学生的统计知识基础相对薄弱，增加了他们在自然语言处理学习方面的困难，甚至会让他们失去兴趣。因此，为了减轻学习者对统计知识的陌生感，本文介绍了一个简单的基于 N 元语言模型的词串生成器。通过这一小程序，学习者能够更加容易地理解统计语言模型的基本运作方式。

3.1 N 元模型

设 S 是一个任意的词序列，即句子，那么 S 可表达为由 n 个词组成的序列，如 $S=w_1w_2\cdots w_n$。那么我们可以设想该词串序列 S 出现的概率 $P(S)$。根据条件概率的链式法则，$P(S)$可表达为：

$$P(w_1w_2\ldots w_n)=\prod_i P(w_i \mid w_1w_2\ldots w_{i-1})$$

就是说，任意的词序列 $w_1w_2\cdots w_n$ 出现的概率是，每个词在它前面所有词出现的条件下出现的条件概率的乘积。可是，这种方法的计算负担太大，根据马尔科夫假设可以做进一步简化。如：

$$P(w_1w_2\ldots w_n)\approx\prod_i P(w_i \mid w_{i-(N\text{-}1)}\ldots w_{i-1})$$

该简化式就是 N 元模型，即假设当前词 w_i 的出现概率只跟它前面的 N-1 个词有关。因此，二元模型（bigram）仅考虑当前词（w_i）的前一个词作为条件概率；三元模型(trigram)则以前两个词的序列作为条件概率。例如：

P (沿着荷塘是一条曲折的小煤屑路)

= p(沿着| *)·p(荷塘|沿着)·p(是|荷塘)·p(一|是) … p(STOP| 路) <二元模型>

= p(沿着| *,*)·p(荷塘|*,沿着)·p(是|沿着,荷塘)… p(STOP| 煤屑,路) <三元模型>

3.2 词串生成器

我们在大规模语料中能够计算出每个词的条件概率分布。这就是统计语言模型。通过我们设计的 N 元词串生成器，学习者可以免除复杂的手工计算过程。在生成器的界面上，使用者首先输入训练语料（1），然后选择二元或三元模型（2），那么词串生成器就会显示任意生成的词串（3）。该词串的生成过程如下：

1）系统任意指定一个词，称为“种子词”

2）根据语言模型的概率分布，选择跟种子词接续概率最高的词，作为下一个词

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3）以步骤 2）中生成的词为下一个种子词，重复步骤 2），循环指定的次数

4）生成的词串显示在窗口（3）

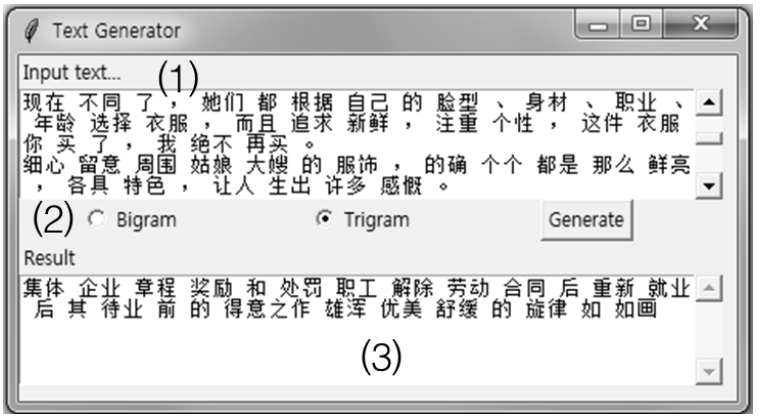

图 2　词串生成器界面

由于该生成器训练出来的概率参数受到输入文本的词或短语的影响，因此学习者可以发现按照训练文本的不同，所生成的词串的内容及文体也不同。同时，教学者可以引导学习者自己体会这种随机生成的文本与真实文本的差异及缺点。这种对比可以让学习者初步认识统计语言模型的优点与缺点。

4　结论

互联网的迅猛发展与普及，引起了人们分析网络自然语言大数据的热潮。由于基于统计的机器学习方法适用于处理以大量、快变、来源领域多样为特点的大数据，因而成为自然语言处理的主流方法。这种思想也深刻地影响到语言学本体研究。以语料库语言学为代表的经验主义研究模式正逐渐扩大语言学领域的外延。随着这种趋势，越来越多的语言研究者及学习者开始关心和采取基于大规模语料的研究路线。可是，对于以文科背景为主的语言学研究者而言，语言模型因要求较多的数学、统计学的基础，往往成为他们难以理解的黑箱子。本文总结了分词及语言模型的要点，并在此基础上构建了面向初学者的中文信息处理平台。

第二节介绍了自动分词/标注器 ICTCLAS，并在此基础上设计了一个用户友好（user-friendly）的 GUI 界面。该界面提供了一个简易的文本处理手段，不熟悉程序语言的使用者也能用以此进行汉语自动分词与词性标注。该程序独立于网络运行，能处理大规模文本，并支持文本编码自动识别及分词结果的储存。第三节的词串生成器能够根据输入文本自动生成出任意词串，以模拟自动句子生成。“自动句子生成”的概念有利于激发学习者的好奇心，词串生成演示流程符合直觉，学习者能够通过使用生成器掌握 N 元模型的基本原理，同时能够体验机器学习的基础概念。

综上所述，本文介绍的平台能满足中文文本分词/词性标注的基本需求，也可服务于初级信息处理教学。我们将在后续研究中对该平台作进一步扩展，增加语块分析、句法分析及文本分类等功能。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参考文献

[1] Bird, Steven, Ewan Klein and Edward Loper. *Natural Language Processing with Python*[M]. O'Reilly Media Inc., 2009.
[2] Manning, Christopher D., and Hinrich Schütze. Foundations of Statistical Natural Language Processing[M]. Cambridge: MIT press, 1999.
[3] Yu, Shiwen et al. 北大语料库加工规范：切分·词性标注·注音[J]. *Journal of Chinese Language and Computing*, 2003 (2): 121-158.
[4] 刘群等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展，2004(8)：1421-1429.
[5] 孙茂松, 邹嘉彦. 汉语自动分词研究评书[J]. 当代语言学，2001(3)：22-32.
[6] 俞士汶等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002(5): 49-64, 2002(6): 58-65.
[7] 俞士汶主编. 计算语言学概论[M]. 北京：商务印书馆，2003.
[8] 詹卫东. 计算语言学与中文信息处理研究近年来的发展综述(2004—2008)[C]. 中国语言学年鉴(2004—2008). 北京：商务印书馆, 待刊.
[9] 张华平. NLPIR/ICTCLAS 2015 分词系统开发文档. http://ictclas.nlpir.org/, 2015.

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net