南京農業大學
NANJING AGRICULTURAL UNIVERSITY

数字人文下的词性自动标注

目录

CONTENT

- 词性自动标注的基本知识
- 词性自动标注在数字人文领域的应用
- 古文词性自动标注的方法

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

词性自动标注的基本知识

1、词性标注

◆词性是依据词的语法功能的分类。词性标注一般指词性的自动标注，就是依据语言学的知识，结合计算机技术，实现对文本中词语词性的标注。

◆对文本中每一个词赋予相应的词性标记，包括对标点符号的标记。它代表了一个词的语法特征，或者叫词类。

◆词性标注样例

实现/v 祖国/n 的/u 完全/a 统一/vn ，/w 是/v 海内外/s

全体/n 中国/ns 人/n 的/u 共同/b 心愿/n 。/w

1、词性标注

◆准确的词性标注是多音字消歧和多义词消歧、机器翻译、信息检索、词典编纂等后续任务的基础工作。

- 确定词的语法功能，为自动句法分析打基础

词性标注：根据词例在词表中寻找所属词型的过程，并且从该词型的潜在语法功能中挑选一个作为正在实现的语法功能

- 在词性标注语料库中检索句法结构

一些句法结构是以特定的词类为标志的，例如介词结构、“把”字结构、方位结构等等。

- 为多音字消歧和多义词消歧提供支持

2、词例、词型和词条

◆词例（word token）是指在言语中出现的具体的词语

◆词型（word type）是指具有相同词汇意义的词例的概括，是语言的词汇中的成员

◆词条（dictionary entry）是指词典中的一个条目，通常把有明显语义关系的词型合为一个词条

◆词型概括必须以词义为依据。同一个词型可以有多个词例。

例：我 经常 花1 钱 买 花2，今天 买 的 这 两 枝 花3 特别 鲜艳。我 自己 没有 种 过 花4。

2、词例、词型和词条

- 我经常花1钱买花2，
- 今天买的这两枝花3特别鲜艳。
- 我自己没有种过花4。

1. 概括为一个词型
2. 花1＝{花1}，花2＝{花2 ,花3 ,花4}
3. 花1＝{花1}，花2＝{花2 ,花3}，花3＝{花4}
4. 有4个词型

- 词型概括必须以词义为依据

2、词例、词型和词条

- 分为两个词条，花1有18个义项，花2只有一个义项（花费），并不表示只有两个词型
- 花1的义项排列：

植物器官；植物；似花之物；烟花；花纹；饰花；多色；眼睛迷乱；衣服磨损；花言巧语；精华……，最后附列作为姓氏的“花”

- 其中有些义项只是语素，未必成词

3、词类体系

◆词类

词类体系是指词类划分的标准，是词性标注的理论基础。一种常见的词类划分标准是依据词具有的语法功能，即它所能占据的语法位置的总和，这意味着词性标注为句法分析服务。

- 实词：名、动、形、数、量、代、副
- 虚词：介、连、助、叹、拟声

3、词类体系

◆语法位置

- 以标志词为参照
  - 例如：受“很”的修饰，能后加“了”。
  - 概括性不足
- 以词类为参照
  - 例如：受程度副词的修饰，能后加动态助词。
  - 循环定义
- 以结构关系为参照
  - 例如：在状中结构中做中心语，在述宾结构中做宾语。
  - 以结构关系为初始知识
- 以结构功能为参照
  - 例如：在体词性结构中做定语，在谓词性结构中做中心语。
  - 以结构功能为初始知识

3、词类体系

例句	语法位置
这本书出版了	在主谓结构中做谓语
出版了这本书	在述宾结构中做述语
辞书出版上取得了很大成绩	在定中结构中做中心语，前面不加“的”
这本书的出版是有原因的	在定中结构中做中心语，前面加“的”
这本书还没列入出版计划	在定中结构中做定语，后面不加“的”
出版的可能性很大	在定中结构中做定语，后面加“的”
这本书的不出版是有原因的	在状中结构中做中心语
……	……

3、词类体系

◆ 分类目的

为（自动）句法分析服务

◆ 理想的情形

先把言语中出现的词例都收集起来，并且一一鉴别其所属的词型，然后按照每个词型的语法功能的总和划分为若干个词类，做成一部**语法词典**

◆ 假定实词的语法功能为N种，每种语法功能的取值为“有”和“无”，那么，N种语法功能最多可区分2N种词类。例如，当N=13时，实词最多分出8192类

3、词类体系

◆ 关于词类划分标准的其他说法

- 形态：对于形态丰富的语言，如俄语、法语等等，这不成问题。汉语缺乏形态变化，因此有人认为汉语的实词无法分类（高名凯）
- 意义：表示人或事物的是名词，表示性质或状态的是形容词，表示动作行为的是动词。
  - “战争”、“手术”表示动作行为
  - “木头桌子”中的“木头”表示桌子的性质
  - “笑着说”中的“笑”表示状态
  - 无益于揭示词的语法功能，也无益于句法分析

3、词类体系

◆ 现有的汉语词类体系

- 自马氏文通开始，汉语词类划分方法都是拿印欧语言的词类体系为蓝本，再结合研究者所观察到的汉语的一些特点做局部的调整，例如增加助词和量词，形成包含名词、代词、动词、形容词、副词、数词、量词、介词、连词、助词、叹词在内的词类体系
- 具体词的归类，则往往参照意义标准
- 后来，北大的词类体系从形容词中分出了区别词和状态词，从助词中分出了语气词等等
- 这种方法虽然简单，但不是以大规模的语法调查为基础的，其适用性值得怀疑

3、词类体系

- 现有词类体系的弊病： 不能真正反映词的语法功能
  - 划归一类的词可能有不同的功能。例如，一般认为名词能做主、宾、定、状、中心语，有条件地做谓语，但许多名词只是有其中某项或某几项功能
  - 不同的词类，功能却基本相同。例如动词和形容词都能做主、宾、谓、定、补、状、中心语
  - 对自动句法分析用处不大，词类序列歧义太复杂
    - 例如V N N的歧义：
    - 动宾＝V+(N+N)，其中N＋N可能是定中或联合或主谓
    - 定中＝(V+N)+N，其中V＋N可能是动宾或定中结构

3、词类体系

◆目前汉语常见的词类体系主要有两种：

●词类多功能说

- 一种词类可以有多种功能，例如动词、形容词和名词，其词类不依出现位置的变化而变化，动词、形容词都可以作谓语、补语，也都可以做主宾语
- 词的语法功能是潜在的，每次只实现它的一种功能，但其它功能并非就消失了
- 汉语的词在实现其语法功能时没有词形变化，因此没有根据说不同位置上的词属于不同词类

3、词类体系

◆目前汉语常见的词类体系主要有两种：

●依句辨品说

□ 应依据词在句子中所实现的功能来确定其词类。动词、形容词在谓语位置上是动词，在主宾语位置上则是名词

这本书出版了	动词
出版了这本书	动词
辞书出版上取得了很大成绩	？名词
这本书的出版是有原因的	名词
这本书还没列入出版计划	？形容词
出版的可能性很大	？形容词
这本书的不出版是有原因的	？？

3、词类体系

◆两种词类体系的对照：

这本书 出版① 了
这本书的出版② 是有原因的
这本书的不出版③ 是有原因的
这本书的暂时不出版④ 是有原因的

●依句辨品说：①是动词，②转化为名词，③和④还原为动词。

●词类多功能说：①～④都是动词。

3、词类体系

◆两种词类体系的兼类情况对照：

●5万词条的词表中，动词约10000个，形容词约2000个

	词类多功能说	依句辨品说
动词兼名词	400	8000～9000
形容词兼名词	30	1700～1800

●根据词类多功能说，动词兼名词、形容词兼名词实际上是一个词条下（或一组同形词中）的几个词型，例如“编辑”、“科学”、“把”；依句辨品说则还把同一个词型的几个词例看做是兼类词。（例如“出版”的例子）

4、词性标记集

◆词性标记集是词类体系在词性标注中的具体体现，这些标记就是词性标注时词类划分的依据。

◆几种有影响力的词性标记集：

- 北京大学计算语言学研究所词性标记集（39个标记）；
- 清华大学计算机系词性标记集（112个标记）；
- 中科院计算所词性标记集（39个标记）；
- 教育部语言文字应用研究所的词性标记集（31个标记）；
- 南京农业大学古汉语词性标记集（21个标记）；
- GB/T20532-2006《信息处理用现代汉语词类标记规范》（49个标记）

4、词性标记集

◆北京大学的词性标记集

Ag	形语素	g	语素	ns	地名	u	助词
a	形容词	h	前接成分	nt	机构团体	Vg	动语素
ad	副形词	i	成语	nz	其他专名	v	动词
an	名形词	j	简称略语	o	拟声词	vd	副动词
b	区别词	k	后接成分	p	介词	vn	名动词
c	连词	l	习用语	q	量词	w	标点符号
Dg	副语素	m	数词	r	代词	x	非语素字
d	副词	Ng	名语素	s	处所词	y	语气词
e	叹词	n	名词	Tg	时语素	z	状态词
f	方位词	nr	人名	t	时间词		

4、词性标记集

◆清华大学的词性标记集

112种标记，其中动词标记15种：

a 形容词	ab 带宾语的形容词	z 状态词
vg 一般动词	va 助动词	vf 形式动词
vi 连系动词	vv 动词之前的趋向动词	
vgj 带兼语的动词	vgs 带小句宾语的动词	
vgv 带动词宾语的动词	vga 带形容词宾语的动词	
vgn 带体词宾语的动词	vgd 带双宾语的动词	
iv 谓词性成语		

4、词性标记集

◆清华语料库中的动词“发展”

有10种标记，除了可能是偶然的标注错误（频率低于10）之外，还有以下8种：

ng（29），vg（1107），
vgb（152），vgn（601），
vgp（464），vgv（14），
vgx（915），vgz（28）

这是依据辨品说的彻底贯彻。

5、兼类词

◆简单来说，就是可划分为多个词类的词。一个词条包含多个词型，且这几个词型的语法功能总和有所区别，这个词条或者跟这个词条写法相同的词例叫做兼类词。

◆对于兼类词的词性标注实际上是一种词性消歧。

5、兼类词

◆兼类词统计数据（一）

从40万字（291623词次）语料统计。

语料标注依北京大学词性标记集。

10813个词型，其中兼类词463个，占4.28%，如“过”兼属助词、动词、副词。

107406词次需要词性消歧，兼类词动态频率：36.8%.

5、兼类词

◆兼类词统计数据（二）

从50万字（287604词次）语料统计

语料标注依清华大学词性标记集。

15920个词型，其中兼类词3559个，占22.4%.

双类词1930个，三类词766个，三类以上的兼类词863个

。

203051词次需要词性消歧，兼类词动态频率：71.6%.

5、兼类词

◆北大词表中词形个数最多的前10种兼类模式

n v（名兼动）	414	b d（区兼副）	44
a v（形兼动）	87	g v（素兼动）	44
d v（副兼动）	57	q v（量兼动）	43
n q（名兼量）	52	a n（形兼名）	31
p v（介兼副）	51	a d（形兼副）	29

有的是造词法的事实，语法演变的事实
有的是同形词的区分或词类体系的问题

6、未登录词

◆与自动分词中的未登录词概念一致，即没有收录在分词词表中，但是需要被正确切分出来的词。

◆对于未登录词的词性标注，可以给定一个开放标记集，如名词、动词，然后把未登录词看成是一个具有全部开放标记的兼类词，从而将词性标注转化为兼类词消歧问题。

词性标注的实质

◆非兼类词

- 直接从词典或训练语料中找出其所对应的词类

◆兼类词

- 根据上下文等语料信息，对多个词类进行消歧

◆未登录词

- 给定一个开放标记集，如名词、动词。
- 把未登录词看成是一个具有全部开放标记的兼类词，从而将词性猜测转化为兼类词消歧问题。
- 构词知识来帮助猜测未登录词的词性。（例如知道“老/小+李”“陈+老师”的形式是名词可以）

词性标注正确率的下限

◆标注方法

- 兼类词选择最高频标记

◆计算公式

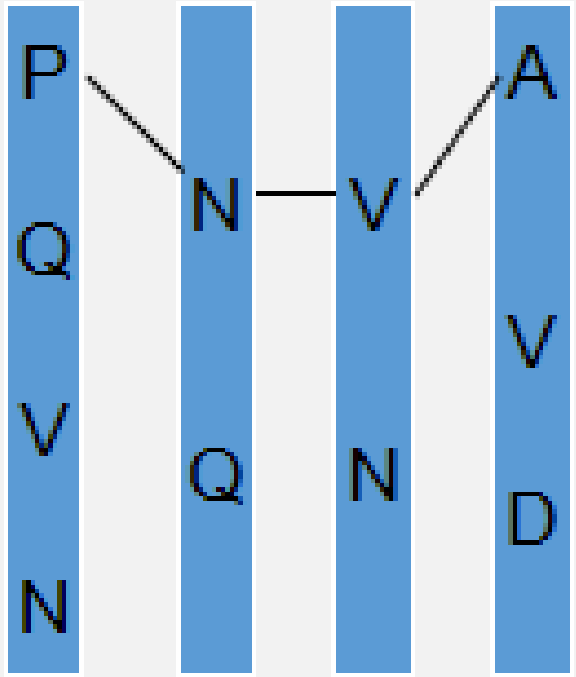
- （单标记词的频率＋兼类词高频标记的频率）/ 总词次

◆50万字语料的计算结果

- 词性标注正确率下限为 85.7%。
- 人名、地名、机构组织名7078词次。
- 假定这些专名都是未登录词，而且词性猜测全部错误，则词性标注正确率的下限将降至83.2%。

哪一种标注最有可能

◆“把 门 锁 好”，4×2×2×3＝48种可能



统计方法就是根据训练集的统计结果来判定哪一种标注概率最大。

对每个标记都选最高频的，这是一种基线标注，常常不能得到整体最优的结果。

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

词性自动标注在数字人文领域的应用

词性自动标注在数字人文领域的应用

◆作为高质量语料库的构成要素

- 语料库中的词是否标注了词性是高质量语料库的一个重要标志。
- 袁悦等探究了不同词性标记集对于典籍实体抽取效果的影响，为词性标记集的选择提供了贡献了理论支撑，对于典籍实体识别工作的改善具有指导作用。
- 留金腾等采用自动分词与词性标注并结合人工校正的方法构建了以《淮南子》为文本的上古汉语分词及词性标注语料库。

词性自动标注在数字人文领域的应用

◆助力汉语言分词与命名实体识别

- 词性标注对于辅助汉语分词与命名实体识别具有积极作用。
- 熊健等提出一种基于分词消歧与词性标注的中文分词方法，该方法首先使用正、逆向最大匹配算法和隐马尔可夫模型完成对文本的分词，得到分词歧义集。然后使用隐马尔可夫模型对文本进行词性标注，词性标注结果用于对分词歧义集进行消歧，该方法有效提高了分词效果。
- 王珊珊等研究了多维领域知识下的《诗经》自动分词，研究表明词性特征的加入，对CRF模型分词性能具有较大影响，从分词结果看，添加了词性特征的模板识别了句中的全部叠词，提高了分词效果。

词性自动标注在数字人文领域的应用

◆为语体风格计算提供支撑

- 宋旭雯以古文版、白话文版、英文版《左传》和《战国策》为例，基于分词和词性标注信息，分析了上述两部典籍不同文体下的语体特征，对词性分布的研究表明，《左传》和《战国策》在名词、动词副词、代词、数词、形容词和量词占比方面具有较高的一致性，然而在虚词的词性占比方面，两部典籍并未保持一致。

- 刘浏基于25部先秦典籍文本的分词和词性标注结果，使用TF-IDF、向量相似度计算、朴素贝叶斯分类器等算法，进行了先秦文本时代特征词的判定研究。

词性自动标注在数字人文领域的应用

◆辅助文本的结构化组织与利用

- 陈诗等将词性信息融入Bi-LSTM-CRF模型用于对典籍的人称指代进行消解，有效提高了人称指代消解的效果。李斌等在对《左传》进行分词与词性标注的基础上，构建了《左传》知识库，并以人物为中心进行了系列的探究。
- 常博林等结合分词与词性标注的方法，构建了《资治通鉴·周秦汉纪》知识库，并搭建了检索系统，提供了含“词性检索”在内的多维度检索入口。

南京農業大學

NANJING AGRICULTURAL UNIVERSITY

古文词性自动标注的方法

1、基于传统机器学习的词性标注方法

◆隐马尔科夫模型（HMM）

- 隐马尔科夫模型（HMM）是自然语言处理（NLP）中一个较为基础的模型，HMM中有两个重要概念——状态和观测值。在词性标注中，词性对应状态，词语序列对应观测值，自动标注的过程就是用观测值预测隐藏状态的过程。

◆CRF模型

- CRF模型可以融入不同的字、词、短语、句子和段落的特征知识，从而有助于所构建相应模型性能的提升。

1、基于传统机器学习的词性标注方法

◆用N元模型做词性标注

- ①根据词语查找候选标记
- ②用动态规划算法求出概率最大的标记串
- 二元模型的两组数据：

  P(t1)：每个标记出现在句首的概率；

  P(ti|ti-1)：每个标记的转移概率。

- 一元模型？

  Baseline

1、基于传统机器学习的词性标注方法

◆N元模型的局限

- 步骤①把一个词语对应多个标记看做是等概率的，这会严重影响标记串构成的正确性。

◆对N元模型加以扩展

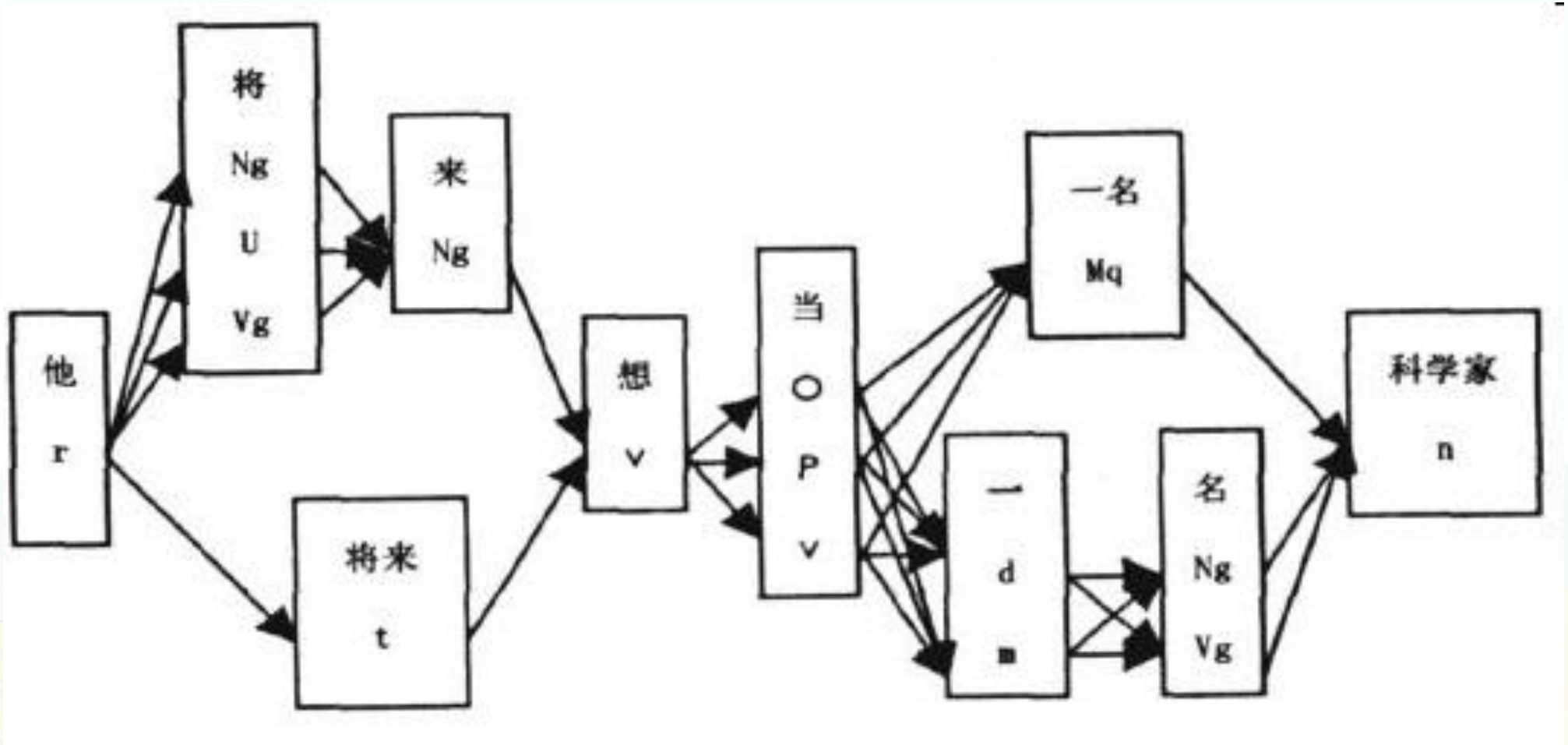
- 用N元模型做词性标注，公式为：$T' = \arg\max_{T} \; p(t_1)\prod_{i=2}^{m} P(t_i \mid t_{i-1})$
- 增加词语的条件概率，公式扩展为：

$$T' = \arg\max_{T} \; p(t_1)\prod_{i=2}^{m} P(t_i \mid t_{i-1})P(w_i \mid t_i)$$

- 扩展后的就是隐马尔科夫模型

1、基于传统机器学习的词性标注方法

◆分词标注一体化（中科院）

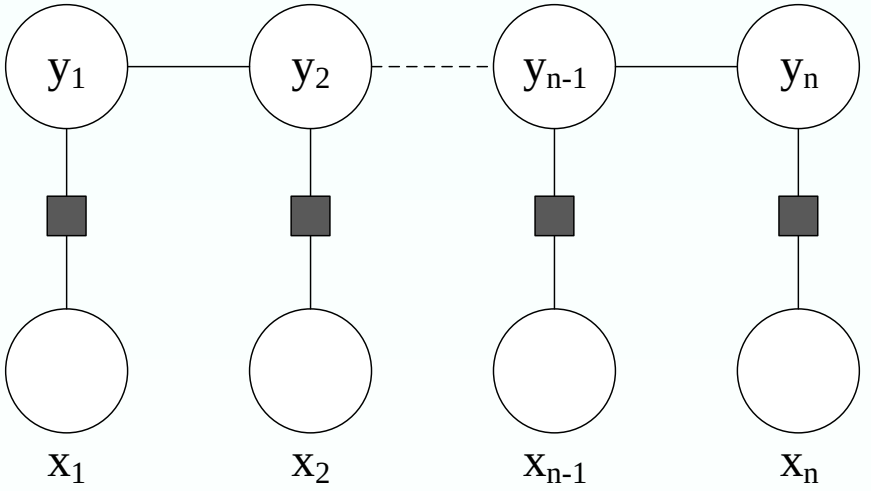


1、基于传统机器学习的词性标注方法--CRF模型

◆CRF是一个无向图模型，具体呈现：

◆线性条件随机场的定义如下：

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp\left\{ \sum_{k=1}^{K} w_k f_k(y_{t-1}, y_t, x_t) \right\}$$



线性链条件随机场模型架构

◆Z（x）为归一化函数

$$Z(x) = \sum_{y} \prod_{t=1}^{T} exp\left\{ \sum_{k=1}^{K} w_k f_k(y_{t-1}, y_t, x_t) \right\}$$

1、基于传统机器学习的词性标注方法--CRF模型

（a）基于CRF的古文词性自动标注语料的预处理

首先确定用于词性标注的词性标记集合。因实验语料为古文语料，故选取面向古文词性标注任务的南京农业大学古汉语词性标记集。原始语料需经过严格分词与词性标注保证数据质量，以获得更好的训练效果，从而为后续研究提供较强的保障。

标注样例：帝/n高陽/nr之/u苗裔/n兮/y，/w朕/r皇考/n曰/v伯庸/nr。/w

1、基于传统机器学习的词性标注方法--CRF模型

（a）基于CRF的古文词性自动标注语料的预处理

将实验语料转换成两列的格式，左边一列为观测序列，即词；右边一列为词对应的词性。以随机的方式将数据集分为十份，其中九份作为训练集，一份为测试集，并使用十折交叉验证的方法，轮流将十份数据其中九份作为训练数据，增强实验的准确性。

观测序帝	Tags	观测序列	Tags
高陽	n	朕	r
之	nr	皇考	n
苗裔	n	曰	v
兮	n	伯庸	nr
，	w	。	w

向	p
世界	n
各国	r
的	u
朋友	n
们	k
，	w

致以	v
诚挚	a
的	u
问候	vn
和	c
良好	a
的	u
祝愿	vn
！	w

1、基于传统机器学习的词性标注方法--CRF模型

（b）编辑模板文件

由于现有的CRF++工具已经封装得较为完善，我们利用CRF进行实验时，可以仅通过修改特征模板template文件以完善实验。

```
# Unigram

U00:%x[-2,0]     # 表示观察当前字与该词前两个字的关系

U01:%x[-1,0]     # 表示观察当前字与该词前一个字的关系

U02:%x[0,0]

U03:%x[1,0]     # 表示观察当前字与该词后一个字的关系

U04:%x[2,0]     # 表示观察当前字与该词后一个字的关系

U05:%x[-1,0]/%x[0,0]

U06:%x[0,0]/%x[1,0]

# Bigram
```

1、基于传统机器学习的词性标注方法--CRF模型

（c）训练、测试与评估命令

训练： crf_learn template train0.txt model1 > time1.log

用template训练出模型model1

测试： crf_test -m model1 test0.txt > output.txt

用model1对测试模型进行序列标注，保存在文件output.txt

评估： python conlleval.py < output.txt

测试训练的模型效果（P、R、F值）

1、基于传统机器学习的词性标注方法--CRF模型

□模型训练命令行参数

-a CRF|MIRA ：算法选择。默认CRF。MIRA算法允许用参数-H 剪枝。

-c float：费用。默认1.0.该值越大，则模型越倾向于训练过度；该值越小，则模型越倾向于训练不足。可通过使用留存数据或者交叉验证来找到最合适的值。

-f int：截断频率。默认为1.若特征的频率小于该值，则会被忽略。对于大语料库，该参数很有用处。

-p int：线程个数。默认为1. 若计算机有多个处理器，该参数可指定线程个数以加快训练过程。

1、基于传统机器学习的词性标注方法--CRF模型

□模型训练命令行参数

-m int：指定最大迭代次数。默认10k。

-e float：设置停止准则。默认0.0001.

-C：将文本格式的模型转换为二进制格式的。

-t：要求得到文本格式的模型。

-H int：该值越小，则剪枝越早发生，可大大减少训练时间，但该值太小则增加复查时间。

1、基于传统机器学习的词性标注方法--CRF模型

□模型训练命令行参数

iter：迭代次数

terr：和tags相关的错误率(错误的tag数/所有tag数)

serr：与sentence相关的错误率（错误的sentence数/所有的sentence数）

obj：当前对象的值。当这个值收敛到一个确定的值时，CRF模型将停止迭代

diff：与上一个对象值之间的相对差

1、基于传统机器学习的词性标注方法--CRF模型

于控制台输入相应指令后，模型即开始训练、测试与评估，待模型训练结束，即可在模型保存路径看到训练好的模型文件，可在控制台看到验证集中词性标注的准确率、召回率与F1值。

```
processed 4983 tokens with 3162 phrases; found: 3062 phrases; correct: 2568.
accuracy:  82.06%; precision:  83.87%; recall:  81.21%; FB1:  82.52
             L/xu: precision: 100.00%; recall: 100.00%; FB1: 100.00  1
               Mg: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
                a: precision:  77.91%; recall:  65.05%; FB1:  70.90  86
               ad: precision:  75.00%; recall:  50.00%; FB1:  60.00  20
               ag: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
               an: precision: 100.00%; recall:  90.00%; FB1:  94.74  9
                b: precision:  92.31%; recall:  63.16%; FB1:  75.00  26
                c: precision:  91.67%; recall:  84.62%; FB1:  88.00  60
               cc: precision:  90.62%; recall:  90.62%; FB1:  90.62  32
                d: precision:  81.70%; recall:  76.22%; FB1:  78.86  153
               dg: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
               dl: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
                f: precision:  92.00%; recall:  88.46%; FB1:  90.20  50
                k: precision: 100.00%; recall:  85.71%; FB1:  92.31  6
```

2、基于深度学习的词性自动标注方法

◆LSTM模型

- LSTM基本结构由输入门、忘记门、输出门三种门结构组成，通过门结构让信息进行选择性地通过，实现所需信息的记忆和其他信息的遗忘。

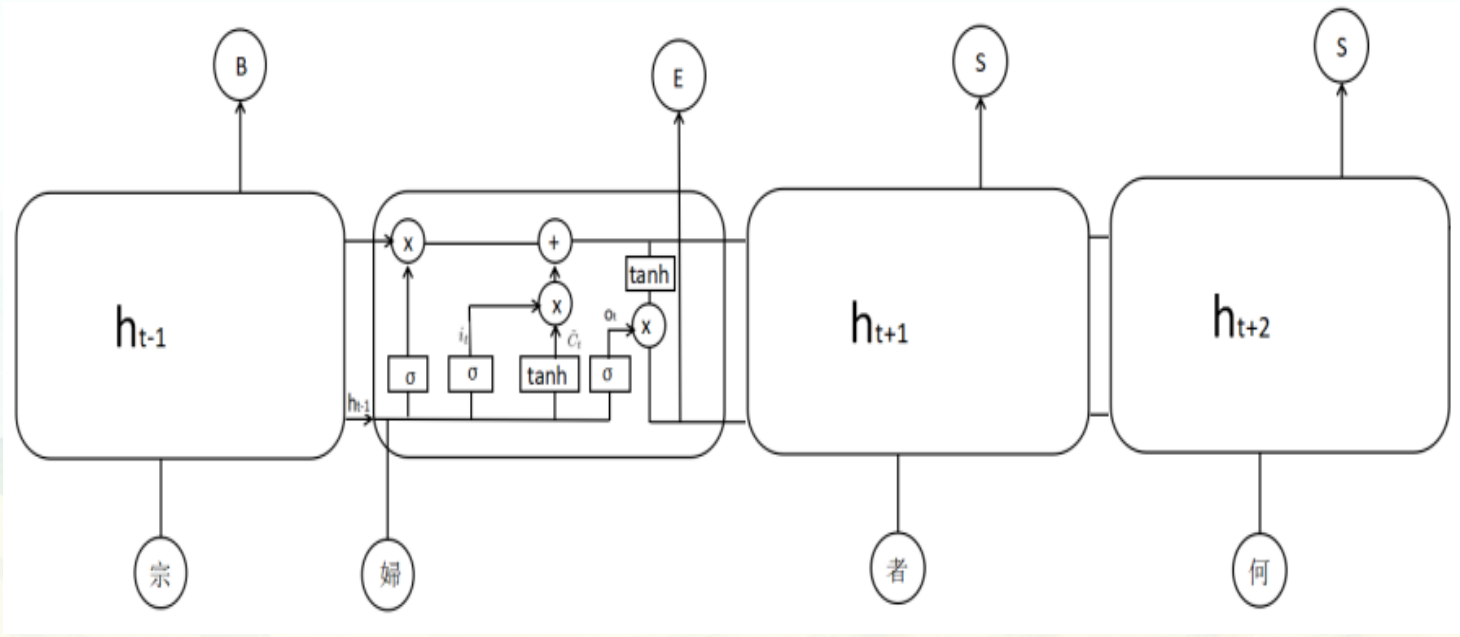
◆BERT模型

- BERT是一种基于Transformer技术的双向编码表示模型，Transformer编码器可以更准确的将人类可读的自然语言转换成机器可识别的形态，从而提高计算机“理解”自然语言的效果。

2、基于深度学习的词性自动标注方法--LSTM模型

（a）模型介绍

LSTM基本结构由输入门、忘记门、输出门三种门结构组成，通过门结构让信息进行选择性地通过，实现所需信息的记忆和其他信息的遗忘。



2、基于深度学习的词性自动标注方法--LSTM模型

（a）模型介绍

不同于RNN结构，隐层中只有简单的单个tanh层，LSTM每个循环模块中有四层结构：3个sigmoid层，1个tanh层。LSTM中还存在其他隐藏状态，一般称之为细胞状态（cell state），记为Ct，呈水平直线贯穿隐藏层，是LSTM的关键环节，线性交互较少，易于保存信息。细胞状态无法选择性的传递信息，更新和保持细胞状态需要借助门结构（gate）来实现，门结构由一个sigmoid层和一个逐点乘积的操作组成。LSTM通过忘记门、输入门、输出门三种门结构实现对细胞信息的增加和删除。

2、基于深度学习的词性自动标注方法--LSTM模型

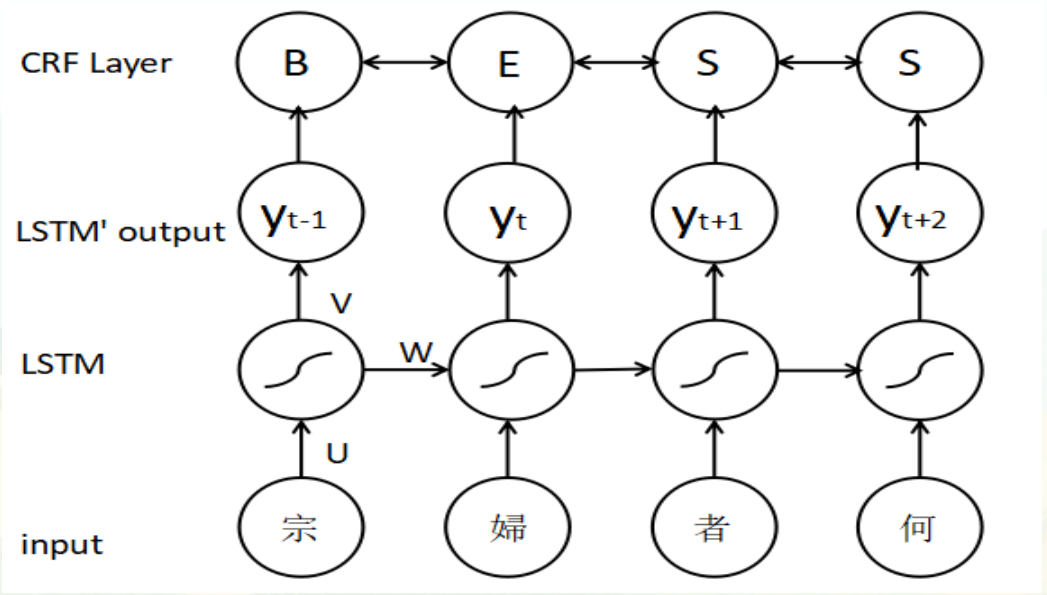
（a）模型介绍

由于在使用LSTM进行实验时，LSTM对数据集的依赖较大，即模型的训练效果与数据集的大小以及质量有很大的关系，并且LSTM不能对未来上下文信息进行分析，例如在给“宗”进行标注时，不能考虑到“婦”及之后的上下文信息，具有局限性。

2、基于深度学习的词性自动标注方法--LSTM模型

（a）模型介绍

CRF使标签标注不仅考虑前文信息，并且受未来状态影响。因此面对具体的任务时，常将CRF引入LSTM中，构成LSTM-CRF模型，即将线性统计模型与神经网络相结合，输出最佳标注序列。LSTM与CRF之间还需要一个转移矩阵A，设P为LSTM的输出矩阵，Ai,j为从i状态转移到j状态的概率，标签序列y的输出为：



$$s(X,y)=\sum_{i=1}^{n}\left(A_{y_i,\,y_{i+1}}+P_{i,\,y_i}\right)$$

2、基于深度学习的词性自动标注方法--LSTM模型

（b）模型关键代码说明

train.py文件：模型训练的主要程序

test.py 文件：测试模型训练效果

config.yml文件：配置文件

2、基于深度学习的词性自动标注方法--LSTM模型

（c）训练指令说明

激活深度学习环境：

```
source activate public_env  # public_env为环境名
```

切换到lstm模型所在文件夹路径：

```
cd lstm_model    # lstm_model为bert模型所在文件夹路径
```

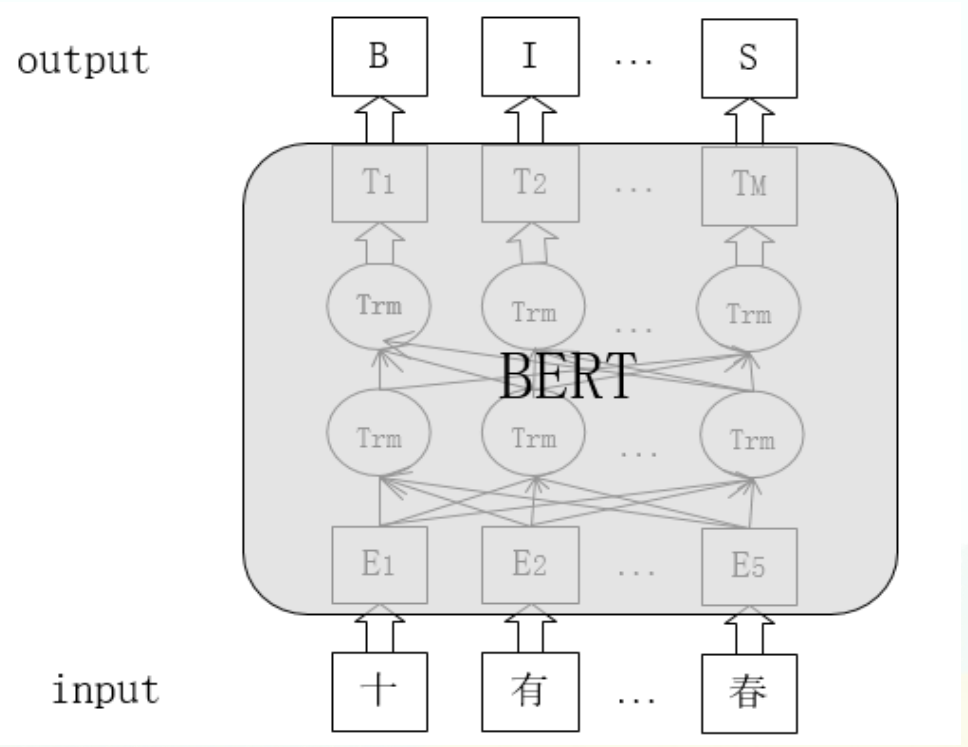
以train.txt为训练集 test.txt为验证集，测试模型效果：

```
python train.py data/train.txt data/test.txt
```

2、基于深度学习的词性自动标注方法--BERT模型

（a）模型介绍

BERT是一种基于Transformer技术的双向编码表示模型，Transformer编码器可以更准确的将人类可读的自然语言转换成机器可识别的形态，从而提高计算机“理解”自然语言的效果。相较于Word2Vec和ELMo，BERT更具优越性，原因在于它通过大量的编码层增强了字嵌入模型的泛化能力，是深层次的双向训练语言模型。



2、基于深度学习的词性自动标注方法--BERT模型

（a）模型介绍

BERT模型在预训练阶段利用Transformer的双向编码器根据上下文双向转换解码。同RNN模型相比，Transformer具有并行化处理功能，为了实现双向理解，使用掩码语言模型（Masked Language Model）遮盖部分词语，并在训练过程中对这些词语进行预测，以及利用下一句预测任务（Next Sentence Prediction），使模型学习两个句子之间的关系。在使用BERT基于大规模语料完成无监督的预训练后，再基于有监督的训练语料对模型进行有监督的微调使其能够应用到各种任务中。

2、基于深度学习的词性自动标注方法--BERT模型

（b）基于BERT模型的古文词性自动标注语料的预处理

- 原始语料需经过严格分词与词性标注保证数据质量，以获得更好的训练效果，从而为后续研究提供较强的保障。
- 基于上述分词和词性标记的结果，将已完成分词和词性标记的语料转换成BERT模型可以识别的格式：

采用四词位标注集{B, M, E, S}给已分词的语料中字符加上标签，其中标签B代表词首字，标签M代表词中间字，标签E代表词末尾字，标签S代表独立成词的单字。

2、基于深度学习的词性自动标注方法--BERT模型

（b）基于BERT模型的古文词性自动标注语料的预处理

- BERT模型在处理文本时是以单个字符为单位的而不是以一个词为单位，这点与使用CRF模型进行词性标注的过程有所区别。
- 该语料预处理方案和模型使用方式事实上可以同时完成分词与词性标注两个任务）。最终实验语料样例如表所示：

观测序列	Tags	观测序列	Tags
帝	S-n	朕	S-r
高	B-nr	皇	B-n
陽	E-nr	考	E-n
之	S-u	曰	S-v
苗	S-n	伯	B-nr
裔	S-n	庸	E-nr
兮	S-y	。	S-w
，	S-w		

2、基于深度学习的词性自动标注方法--BERT模型

（c）标签设置与模型参数设置

- 标签设置

按自己的需求修改setting.py里的类别，确保每个训练标签均存储于setting.py文件内

```
LABELS=["X","O",'B-nr', 'M-nr', 'E-nr', 'S-nr', 'B-n', 'M-n', 'E-n', 'S-n', 'B-w', 'M-w', 'E-w', 'S-w',
'B-ns', 'M-ns', 'E-ns', 'S-ns', 'B-u', 'M-u', 'E-u', 'S-u', 'B-v', 'M-v', 'E-v', 'S-v', 'B-p', 'M-p', 'E-p', 'S-p',
'B-nx', 'M-nx', 'E-nx', 'S-nx', 'B-d', 'M-d', 'E-d', 'S-d', 'B-r', 'M-r', 'E-r', 'S-r', 'B-a', 'M-a', 'E-a', 'S-a',
'B-c', 'M-c', 'E-c', 'S-c', 'B-t', 'M-t', 'E-t', 'S-t', 'B-m', 'M-m', 'E-m', 'S-m', 'B-q', 'M-q', 'E-q', 'S-q',
'B-y', 'M-y', 'E-y', 'S-y', 'B-j', 'M-j', 'E-j', 'S-j', 'B-nc', 'M-nc', 'E-nc', 'S-nc', 'B-nrx', 'M-nrx', 'E-nrx',
'S-nrx', 'B-f', 'M-f', 'E-f', 'S-f', 'B-gv', 'M-gv', 'E-gv', 'S-gv', 'B-i', 'M-i',
'E-i','S-','S-i","[CLS]","[SEP]"]
# [CLS]与[SEP]是模型需要的有特殊作用的标志位，需保留，无需修改。
```

2、基于深度学习的词性自动标注方法--BERT模型

（c）标签设置与模型参数设置

- 模型参数设置

```
CUDA_VISIBLE_DEVICES=1 # 指定训练GPU
nohup # 不挂断运行，以便在注销后命令在后台继续运行
python run_ner.py # 运行训练程序
data_dir=train_data/data_0/ \  # 存放训练语料的文件夹路径
--bert_model=pretrain_models/siku_roberta/ \  # 预训练模型选择
--task_name=ner \  # 任务名，即定义的数据处理类的键
```

2、基于深度学习的词性自动标注方法--BERT模型

（c）标签设置与模型参数设置

- 模型参数设置

```
--output_dir=output/data \    # 存放输出文件的路径
--max_seq_length=128 \   # 最大输入序列长度
--do_train  --eval_batch_size=64  --train_batch_size=64
# 每批次训练数据量大小
--num_train_epochs 10 \    # 迭代次数
--do_eval --warmup_proportion=0.4   # 预热学习率
>  logout.log  2>&1  &  echo  $!     #  输出日志信息到文件
logout.log
```

2、基于深度学习的词性自动标注方法--BERT模型

（d）训练指令与实验结果评估

□ 训练指令

训练指令：

```
source activate public_env  # public_env为环境名
```

切换到模型所在文件夹路径：

```
cd bert_model     # bert_model为bert模型所在文件夹路径
```

打开命令提示符输入指令：

```
sh run.sh
```

2、基于深度学习的词性自动标注方法--BERT模型

（d）训练指令与实验结果评估

□ 训练指令

模型即开始训练，可在生成的日志文件logout.log中看到实时损失（loss）、训练进度的变化情况。

待模型训练结束，可在模型保存路径看到训练好的模型文件，在run_ner.py文件中提供的方法是先运行完所有的epochs之后，再加载模型进行验证。我们在logout.log文件中能看见模型训练的评估结果。

```
Iteration:   5%|▌         | 16/328 [00:16<05:28,  1.05s/it]▯[A▯[A
Iteration:   5%|▌         | 17/328 [00:17<05:27,  1.05s/it]▯[A▯[A
Iteration:   5%|▌         | 18/328 [00:18<05:26,  1.05s/it]▯[A▯[A
Iteration:   6%|▌         | 19/328 [00:19<05:25,  1.05s/it]▯[A▯[A
Iteration:   6%|▌         | 20/328 [00:21<05:24,  1.05s/it]▯[A▯[A
Iteration:   6%|▋         | 21/328 [00:22<05:23,  1.05s/it]▯[A▯[A
Iteration:   7%|▋         | 22/328 [00:23<05:22,  1.05s/it]▯[A▯[A
Iteration:   7%|▋         | 23/328 [00:24<05:21,  1.05s/it]▯[A▯[A
Iteration:   7%|▋         | 24/328 [00:25<05:20,  1.05s/it]▯[A▯[A
```

2、基于深度学习的词性自动标注方法--BERT模型

（d）训练指令与实验结果评估

□ 实验结果评估

依然采用准确率P（Precision）、召回率 R（Recall）和调和平均值 F（F-measure）作为模型分词效果的评测指标

```
Evaluating: 100%|████████████████████| 37/37 [00:11<00:00,  3.90it/s][
06/15/2021 12:14:08 - INFO - __main__ -
              precision    recall  f1-score   support

           p     0.9375    0.9652    0.9512      1150
           n     0.7857    0.7791    0.7824      7152
           w     0.9969    0.9970    0.9969      8636
           v     0.8870    0.8791    0.8830      8965
           y     0.9692    0.9829    0.9760      1759
           d     0.8842    0.9377    0.9102      2053
           c     0.7276    0.9362    0.8188       987
           r     0.9412    0.9613    0.9512      3232
          nr     0.7752    0.7442    0.7594      1372
           u     0.9621    0.9253    0.9434      1098
           m     0.9062    0.9333    0.9196       435
```

3、基于古汉语典籍词性自动标注的数字人文研究

◆以二十四史文本为语料，利用SIKU-BERT典籍智能处理系统的“语料库模式”进行基于词性自动标注数字人文研究。

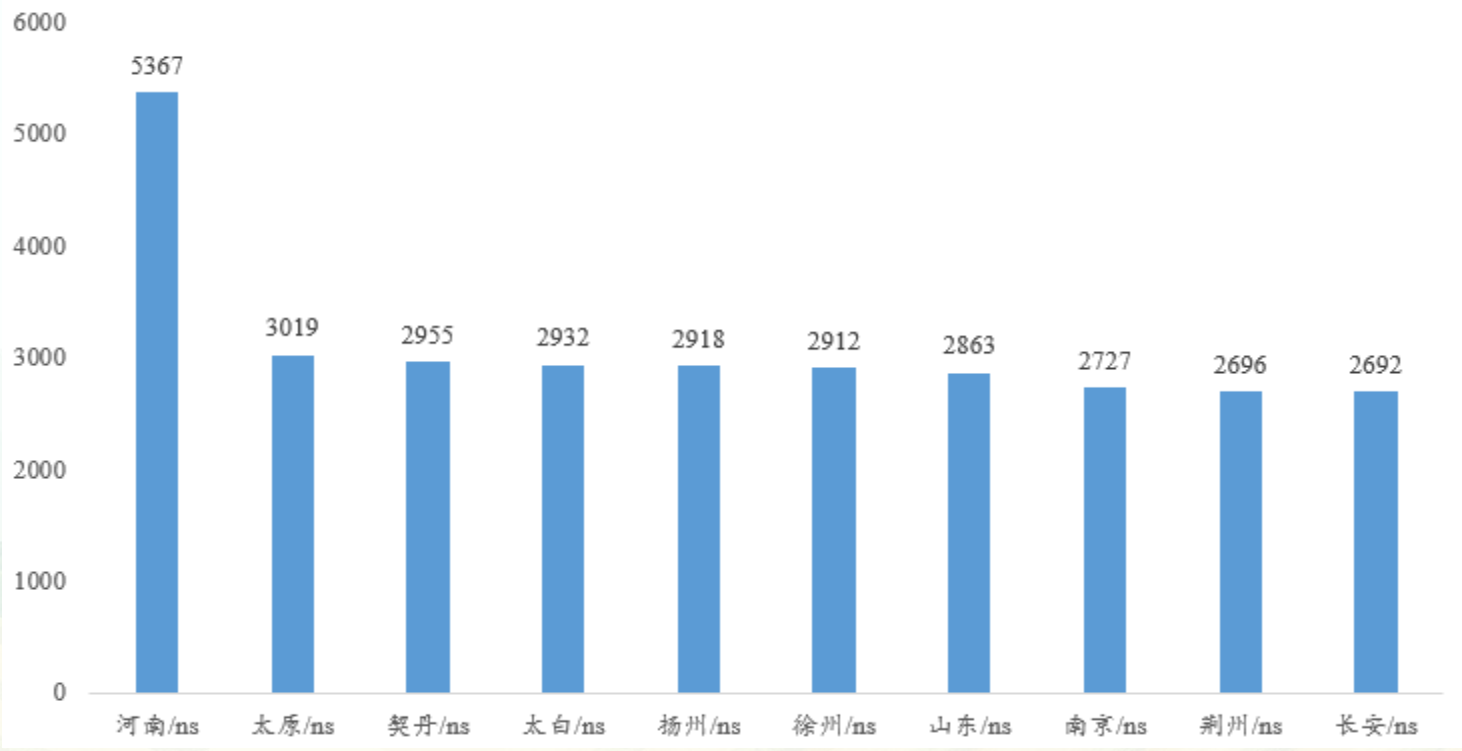
◆名词词性标注的频次结果

词性	频次	释义	举例
n	4039664	一般名词	鬼神、山川
nh	592986	人名	轩辕
ns	443094	地名	襄平县
nt	353532	时间名词	春、夏、秋、冬
nd	300855	方向名词	东、西、南、北
nl	45351	地点名词	城北
nz	33762	其他专有名词	山海经
ni	572	机构名称	辽队

3、基于古汉语典籍词性自动标注的数字人文研究

◆地名词语频率分布

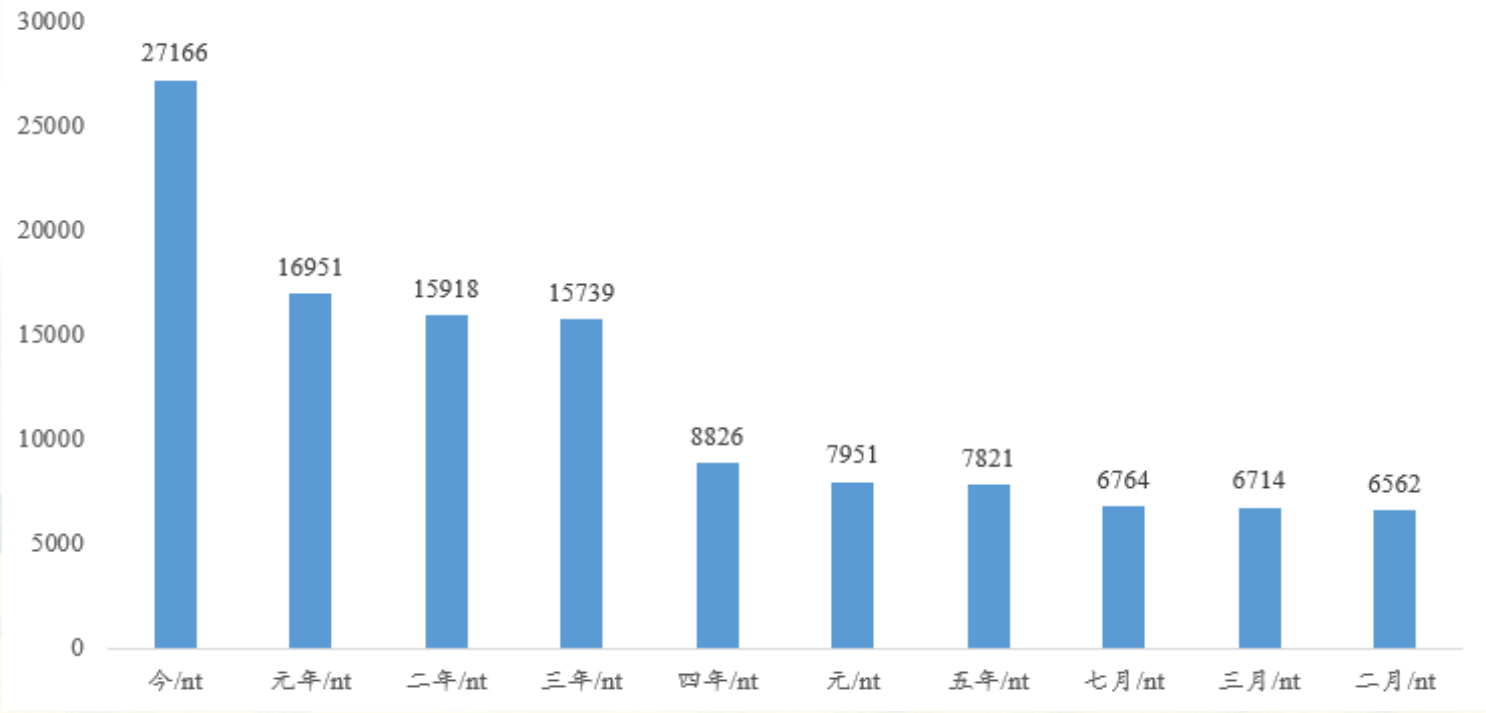
通过频次分析可知哪些地域历来为兵家必争之地。频次排在首位的“河南”非指今日中国的省份,而是多指古代河套以南地区。



3、基于古汉语典籍词性自动标注的数字人文研究

◆时间词语频率分布

通过频次分析可知历史上权力更迭与事件频发的时间段，从而开展更为深入的史学知识挖掘与分析。



3、基于古汉语典籍词性自动标注的数字人文研究

◆时间词语频率分布

从“元年”“二年”“三年”“四年”之类的时间名词可知, 王朝更替或权力更迭初期往往发生重要历史事件。更为有趣的是, “七月”“三月”“二月”三个月份也是历史上事件多发时间段, 个中规律值得跨学科合作下的深度挖掘。

本章结束