汉语分词与标注

汇报人：山巾芝 朱超杰 余路遥 李乐凡 蒋凌昀 刘家昌

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目 录

CONTENTS

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

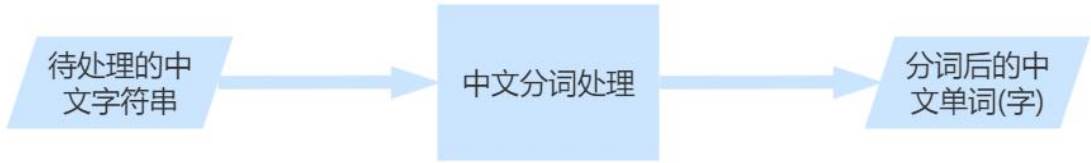
1

经典综述

主讲人：山巾芝 朱超杰

汉语分词

中文分词指的是中文在基本文法上有其特殊性而存在的分词，简而言之，就是将一句话切分成一个个单词的过程。



我是北京理工大学学生 ➡ 我 | 是 | 北京理工大学 | 学生

不同的分词算法可能会得到不同的分词结果

为什么要汉语分词

- 英文以空格作为天然的分隔符，而中文词语之间没有分隔。

  I major in English. ⟶ 我的专业是英语。

- 在中文里，“词”和“词组”边界模糊。

  对 随 地 吐 痰 者 给 予 处 罚。

- 后续工作：汉字处理、信息检索、内容分析、语音处理等。

  雅虎中国网页搜索部总监张勤认为，中文分词是搜索技术的基础，只有做好了分词，才能有好的搜索。

汉语分词难点——切分歧义

- 交集型切分歧义OAS（交叉歧义）—— 对于汉字串AJB，AJ、JB同时成词

  “人民生活幸福” → ① 人民 | 生活 | 幸福
  ② 人 | 民生 | 活 | 幸福

- 组合型切分歧义CAS（覆盖歧义）—— 对于汉字串AB，A、B、AB同时成词

  “校友会” → ① 我 | 在 | 校友会 | 工作
  ② 我 | 的 | 校友 | 会 | 来

- 真歧义 —— 本身的语法和语义都没有问题，即便人工进行切分也会产生歧义

  “乒乓球拍卖完” → ① 乒乓 | 球拍 | 卖完
  ② 乒乓球 | 拍卖 | 完

汉语分词难点——未登录词识别

未登录词有两种，一种指已有的词表中没有收录的词，另一种指训练语料中未曾出现过的词，而后一种也可被称作集外词（Out of Vocabulary，OOV），即训练集以外的词。

未登录词通常包含以下几种类型：

- 新出现的普通词汇。如网络用语中层出不穷的新词等。
- 专有名词。如人名、地名以及组织机构名称等。
- 专业名词和研究领域名称。将分词运用到某些特定领域或专业，需要特定的领域词典。
- 其它专用名词。如新产生的产品名、电影名称、书籍名称等。

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

汉语分词发展历程

俄汉翻译机的研制时期，苏联研究汉俄机器翻译的学者首先提出的、后来被称为6-5-4-3-2-1查词法。

上世纪50年代后期

Sproat 等首次**基于统计学习方法**实现中文分词。根据处理的粒度，分为基于词和基于字两类标注。

1990年

Collobert 等首次将**深度学习算法**引入自然语言任务中。该方法可以通过最终的分词标注训练集，有效学习原始特征和上下文表示。

2011年

随后CNN、GRN、LSTM、BiLSTM等深度学习模型都被引入到中文分词任务中，并结合中文分词进行多种改进。

汉语分词算法分类

- 基于词典的分词算法
- 统计机器学习的算法
- 基于神经网络的算法

基于词典的分词算法（也称机械分词）

按照扫描方向的不同，串匹配分词方法可以分为正向匹配和逆向匹配。

按照不同长度优先匹配的情况，可以分为最大(最长)匹配和最小(最短)匹配。

常用的几种机械分词方法如下：

- 正向最大匹配算法FMM
- 逆向最大匹配算法RMM
- 双向最大匹配算法BM

基于词典的分词算法（也称机械分词）

- 正向最大匹配算法FMM（从左到右的方向）

  ⟶ 永 和服 装 有限公司

- 逆向最大匹配算法RMM（从右到左的方向）

  ⟶ 永 和 服装 有限公司

词　　典：服装 有限公司 公司 和服

待分词序列：永和服装有限公司

统计结果表明，单纯使用正向最大匹配的错误率为1/169，单纯使用逆向最大匹配的错误率为1/245，显然RMM法在切分的准确率上比FMM法有很大提高。但这种精度还远远不能满足实际的需要。实际使用的分词系统，都是把机械分词作为一种初分手段，还需通过利用各种其它的语言信息来进一步提高切分的准确率。

基于词典的分词算法（也称机械分词）

- 双向最大匹配算法BM
  - 比较FMM和RMM得到的分词结果，如果两种结果相同，则认为分词正确，否则，按最小集处理。
  - 在实用中文信息处理系统中得以广泛使用的原因

90%左右的句子	9%的句子	不到1%的句子
切分结果完全重合且正确	切分结果不同， 但其中必有1个是正确的	切分结果重合却是错误的 或者不重合但两个都是错误的

基于词典的分词算法（也称机械分词）

词典分词方法包含两个核心内容：分词算法与词典结构。

- 算法设计可从以下几方面展开
  - 字典结构改进
  - 改进扫描方式
  - 将词典中的词按由长到短递减顺序逐字搜索整个待处理材料，直到分出全部词为止
- 影响词典性能的三个因素
  - 词查询速度
  - 词典空间利用率
  - 词典维护性能（e.g. 设计Hash表）

基于词典的分词算法（也称机械分词）

- 优点：
  - 易于实现
  - 可以精确地切分出所有在词典中存在的词
- 缺点：
  - 匹配速度慢
  - 存在交集型和组合型歧义切分问题
  - 词本身没有一个标准的定义，没有统一标准的词集
  - 不同词典产生的歧义也不同

统计机器学习的算法

- 主要思想

  把每个词看作字组成，相邻的字在语料库中出现的次数越多，就越可能是一个词。

- 主要模型
  - N-gram模型
  - 最大熵模型ME
  - 隐马尔可夫模型HMM

N-gram模型

- 主要思想

  第n个词的出现只与前面n-1个词相关，与其他词都不相关，整个语句的概率就是各个词出现概率的乘积。

- 算法推论

  假设一个字符串s由m个词组成，因此我们需要计算出P(w1,w2,… ,wm)的概率，根据概率论中的链式法则得到如下：

  $P(w1,w2,...,wm) = P(w1)*P(w2|w1)*P(w3|w1,w2|)...P(wm|w1,w2...wm-1)$

  根据马尔科夫假设，当前词仅与前面几个词相关，所以不必追溯到最开始的那个词，即 $P(wi|w1,w2...wi-1) = P(wi|wi-n+1,wi-1)$，

  $P(s) = P(w1,w2,...,wm) = P(w1)*P(w2|w1)*P(w3|w1,w2|)...P(wm|w1,w2...wm-1) \approx P(w1)*P(w2|w1)*P(w3|w2)...P(wm|wm-1)$

最大熵模型ME

- 最大熵原理

  对一个随机事件的概率分布进行预测时，预测应当满足全部已知的约束，而对未知的情况不要做任何主观假设。在这种情况下，概率分布最均匀，预测的风险最小，因此得到的概率分布的熵是最大。

  一个朴素的说法：不要把所有的鸡蛋放在一个篮子里。

- 主要思想

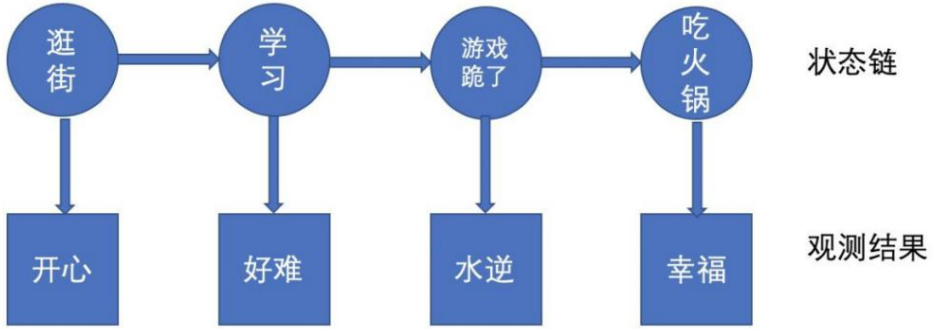
  在学习概率模型时，所有可能的模型中熵最大的模型是最好的模型；若概率模型需要满足一些约束，则最大熵原理就是在满足已知约束的条件集合中选择熵最大模型。

最大熵模型ME

- 优点
  - 最大熵统计模型获得的是所有满足约束条件的模型中信息熵极大的模型，作为经典的分类模型时准确率较高。
  - 可以灵活地设置约束条件，通过约束条件的多少可以调节模型对未知数据的适应度和对已知数据的拟合程度。
- 缺点
  - 由于约束函数数量和样本数目有关系，导致迭代过程计算量巨大，实际应用起来比较难。

隐马尔可夫模型HMM

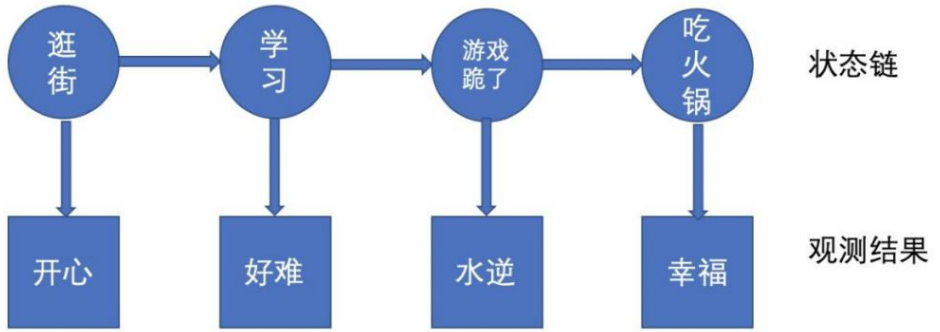
- 隐马尔可夫模型是关于时序的概率模型。描述由一个隐藏的马尔可夫链随机生成的不可观测的状态序列，再由各个状态生成一个观测从而产生观测序列的过程。



- 五元组
  - I：状态序列，如 I={学习，逛街，学习，游戏，吃火锅}
  - O：观测序列，如O={难顶，开心，很烦，心累，幸福}
  - A：状态转移概率矩阵，表示从t时刻状态qi，t+1变成qj的概率，
    如前一天逛街变换到今天学习的概率
  - B：观测转移概率矩阵，表示从t时刻状态qj，产生观测结果vk的概率，
    如今天逛街然后今天是开心的概率
  - π：初始状态概率分布，表示在t=1时刻处于状态qi的概率，
    如第一天是逛街的概率

$\lambda = (A, B, \pi)$

隐马尔可夫模型HMM



- 两个假设
  - 齐次马尔可夫链假设：任一时刻的状态只与上一时刻的状态有关，与其他时刻的状态、观测无关。公式描述：$P(i_{t+1}|i_1,i_2\ldots i_t;o_1,o_2\ldots..o_t)=P(i_{t+1}|i_t)$ 。

    【明天所做的事情只与今天所做的有关，比如今天逛街（状态，这是你看不到的）那么明天很有可能就是学习（因为昨天浪了一天），但是与昨天学习无关，与每天的心情（观测，这是你在朋友圈看到的）也无关。】
  - 观测独立性假设：任一时刻的观测只与当前时刻的状态有关。

    公式描述：$P(o_t|i_1,i_2\ldots i_T;o_1,o_2\ldots..o_{t-1},o_{t+1}\ldots o_T)=P(o_t|i_t)$ 。

    【今天的心情（观测）只与今天所做的事有关（状态）。如我们今天感觉倒霉（观测）因为今天晋级赛跪了。而与昨天辛苦工作（状态），明天还要辛苦工作无关。】

隐马尔可夫模型HMM

- 三个问题
  - 概率计算问题：给定模型λ=(A,B,π)和观测序列O，计算在模型λ下观测序列出现的最大概率P(O|λ)。(Forward-backward算法)
  - 学习问题：给定观测序列O，计算模型的参数λ，使得在该参数下观测序列出现的概率最大，即P(O|λ)最大。（Baum-Welch算法)
  - 预测问题：我们已经获取了模型λ=(A,B,π)和观测序列O，计算最有可能的状态序列I。（Viterbi算法)

统计机器学习的算法

- 优点：较好地识别未登录词和消除歧义
- 缺点：统计模型复杂度高，运行周期长，依赖人工特征提取
- 随着计算机运行速度加快，神经网络逐渐进入分词领域

基于神经网络的算法

- 该方法是模拟人脑并行，分布处理和建立数值计算模型工作的。 它将分词知识所分散隐式的方法存入神经网络内部，通过自学习和训练修改内部权值，以达到正确的分词结果，最后给出神经网络自动分词结果。
- 自序列标注方法在 bakeoff 测试中取得优异成绩后，将神经网络与序列标注相结合成为中文分词领域的通用框架。
- 主要模型
  - 循环神经网络RNN
  - 长短期记忆人工神经网络LSTM
  - 门控循环单元GRU

词性标注

为每一个词的词性加上标注。也就是确定该词属于名词、动词、形容词还是其他词性的过程。

对于几乎所有的语言处理任务来说，词性标注都是很重要的前置处理任务。

他/r 做/v 了/u 一/m 个/q 报告/n

在分词中的应用

影响分词效果的主要问题

- 未登录词（主）
- 歧义

Bakeoff-2003的评测

参赛队	召回率 R	精确率 P	调和均值 F	R_{OOV}	R_{IV}
UC Berkley	0.966	0.956	0.961	0.364	0.980
Nianwen Xue	0.961	0.958	0.959	0.729	0.966
Nara IST Japan	0.944	0.945	0.945	0.574	0.952

在整体质量较高的情况下，基于字标注的分词系统有着明显的未登录词识别优势

分类

基于规则

利用现有的语言学成果，总结出有用的规则。在基本标注的情况下，结合上下文和规则库消除歧义，保留唯一合适的词性。

基于统计

对于给定的输入词串，先确定所有可能的词性串，选出得分最高的作为最佳输出。

基于深度学习

依靠神经网络强大的特征提取和表征能力来进行文本数据的处理。

基于规则的词性标注

由词性标注的规则组成的规则库

”ADJ”+”NUM”	形容词+数词
” V”+”ADJ”	动词+形容词
”V”+”PRON”	动词+代词

基于规则的词性标注

早期，词性标注的规则库需要人工构造，艰难耗时。基于转换的错误驱动的方法首次克服了手工制定规则的问题。

01

1971年，TAGGIT系统被用于Brown语料库的辅助词性标注工作。

02

1995年，Eric Brill提出了基于转换的错误驱动的方法。

03

2000年，李晓黎等人提出用数据采掘的方法获取汉语词性标注规则。

04

2008年，王广正等人提出了基于规则优先级的词性标注方法。

基于统计的词性标注

- 隐马尔可夫模型（HMM）
- 最大熵模型
- 条件随机场（CRF）

条件随机场（CRF）

特征函数

$$f_1(s, i, l_i, l_{i-1}) = 1$$

对于句子s，标注l对第i位的标注为l_i的情况下，满足l_{i-1}函数值为1，否则为0。

特征函数集

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

若f为1表示特征合理，λ为正，否则为负

基于深度学习的词性标注

词嵌入

- 独热编码（One-hot)

{1，0，0，0}，{0，1，0，0}，{0，0，1，0}，{0，0，0，1}

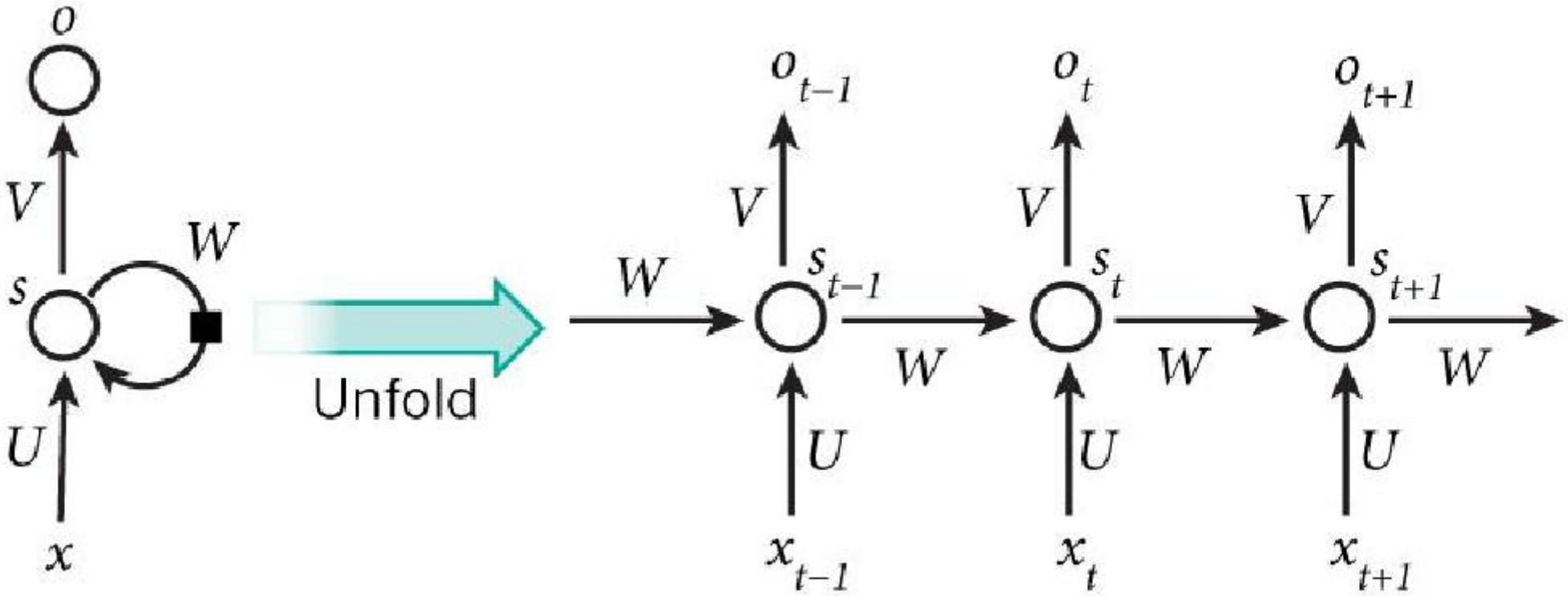
- Word2Vec模型

利用局部上下文。低维，稠密

- GloVe(Global Vectors for Word Representation)

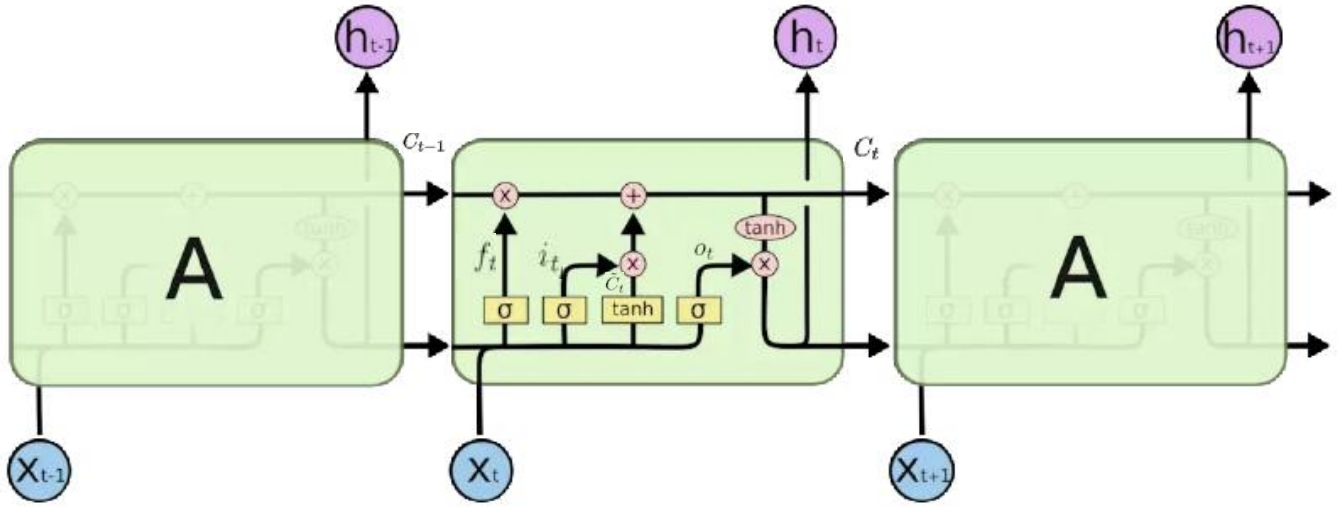
增加了对全局特征的利用。

循环神经网络（RNN）



针对序列信息进行特征抽取

长短期记忆人工神经网络（LSTM）



记住需要长时间记忆的，忘记不重要的信息

遗忘门 $f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$

输入门 $i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$

输入值 $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$

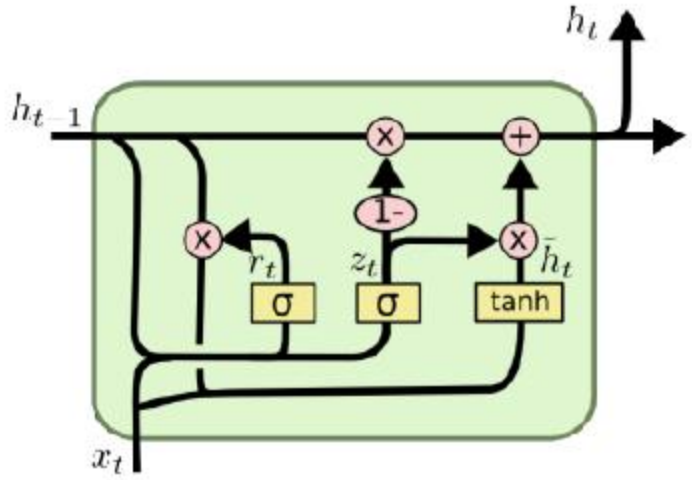
$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

输出门 $o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$

$h_t = o_t * \tanh\left(C_t\right)$

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

门控循环单元（GRU）



更新门 $z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$

重置门 $r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$

候选状态 $\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

相较于LSTM，减少了一个“阀门”，
单元结构更加简单，性能更强

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

2

分词与标注联合模型

主讲人：余路遥

分词与标注联合模型

流水线模型：先分词，再标注

联合模型：分词与标注同时进行

优势

流水线模型会导致错误的传播，并且分词的结果对标注的效果影响很大，同时标注也能为分词分词任务提供重要信息，因此使用联合模型。

实现思路

- 序列标注模型
- Transition-based系统

序列标注模型

将分词任务拓展为分词标注任务：将边界标签拓展到加入词性标签：

狗/S 蹲/S 在/S 墙/B 角/E

↓

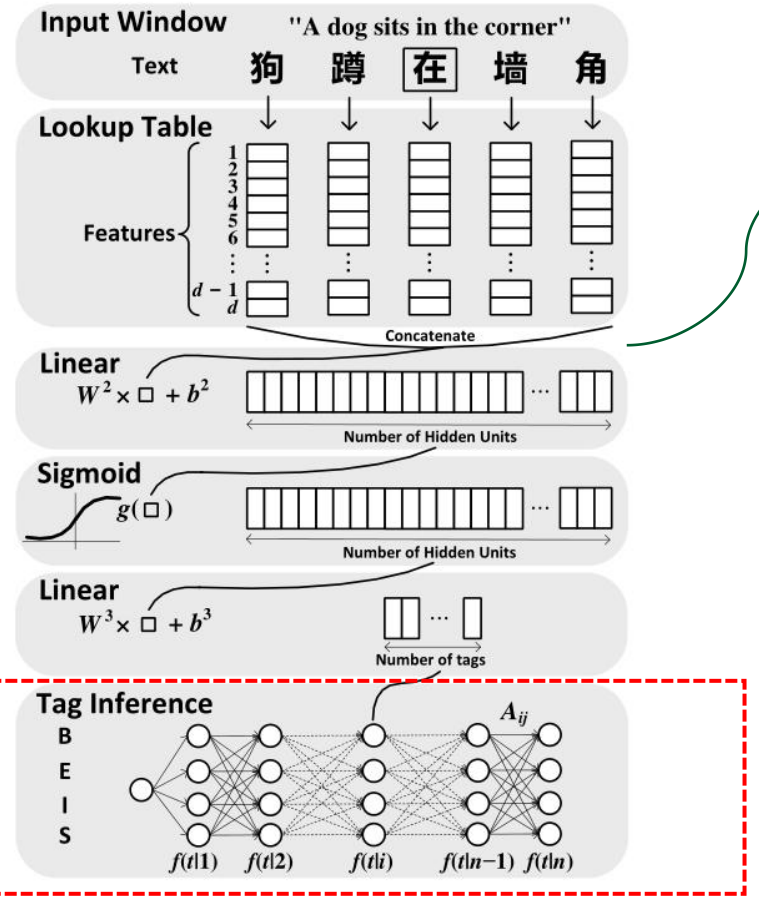
狗/S_NR 蹲/S_VP 在/S_P 墙/B_NR 角/E_NR

传统统计学习的方法：需要大量的特征
1.模型过大难以存储和计算；
2.参数过多而造成过拟合；
3.耗费计算时间；4.难以解码

→

深度学习的方法：
使用神经网络来
直接获取单词之间的更高层的特征表示

序列标注模型

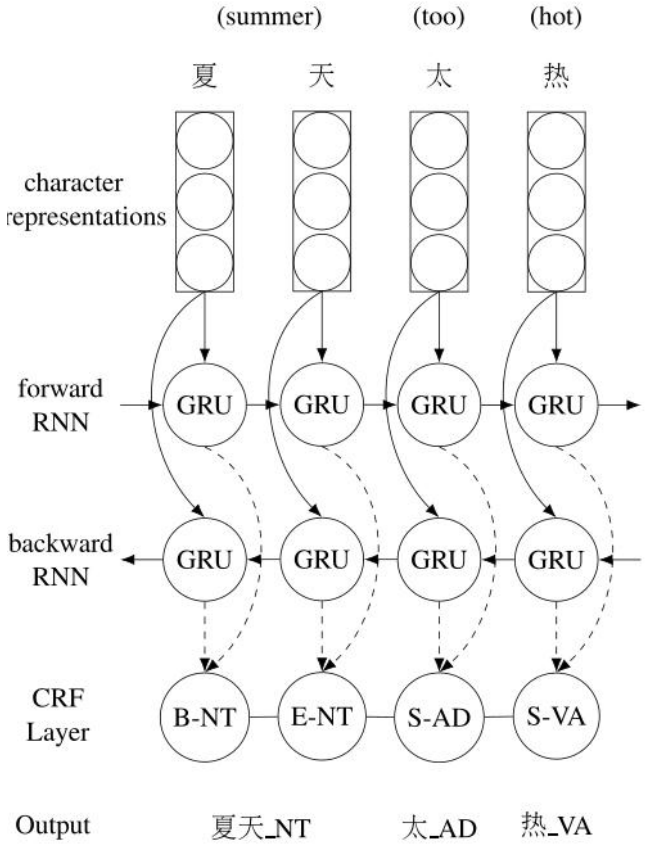


$$f_\theta^1(c_i) = \begin{pmatrix} \mathcal{Z}_\mathcal{D}(c_{i-w/2}) \\ \vdots \\ \mathcal{Z}_\mathcal{D}(c_i) \\ \vdots \\ \mathcal{Z}_\mathcal{D}(c_{i+w/2}) \end{pmatrix}$$

初始化特征向量，经过神经网络获得一个句子中每个位置上的单词分别对应标签集中每个标注的得分。
通过维比特算法，找到该句子的最优的标签序列。

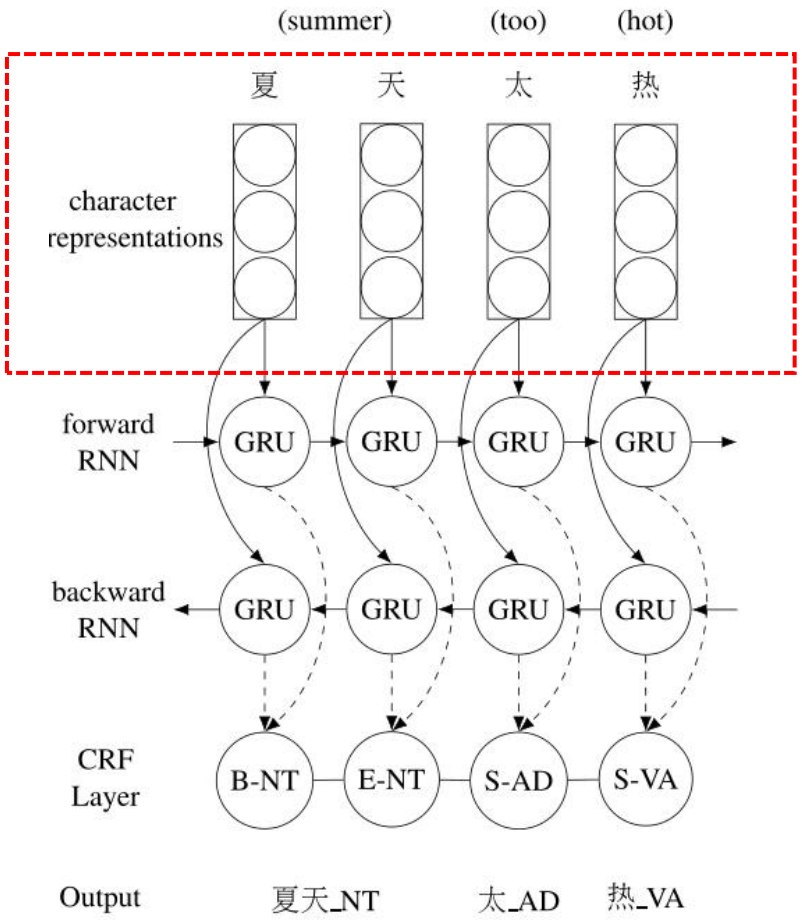
序列标注模型

双向的RNN来获取双向的特征表示并对标签进行预测，最后经过CRF来选择出合适的标签序列。

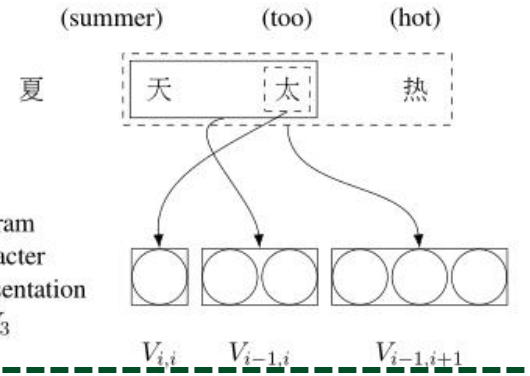


北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

序列标注模型



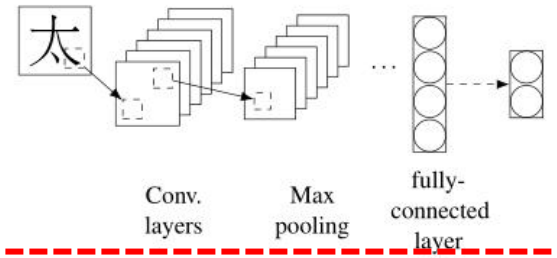
Concatenated N-gram



Radicals Features

获得字的偏旁：“钅”：银、铝、铁

Orthographical Feature



Pre-trained Character Embeddings

利用Wikipedia和SogouCS语料训练字符的GloVe向量

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

transition-based系统

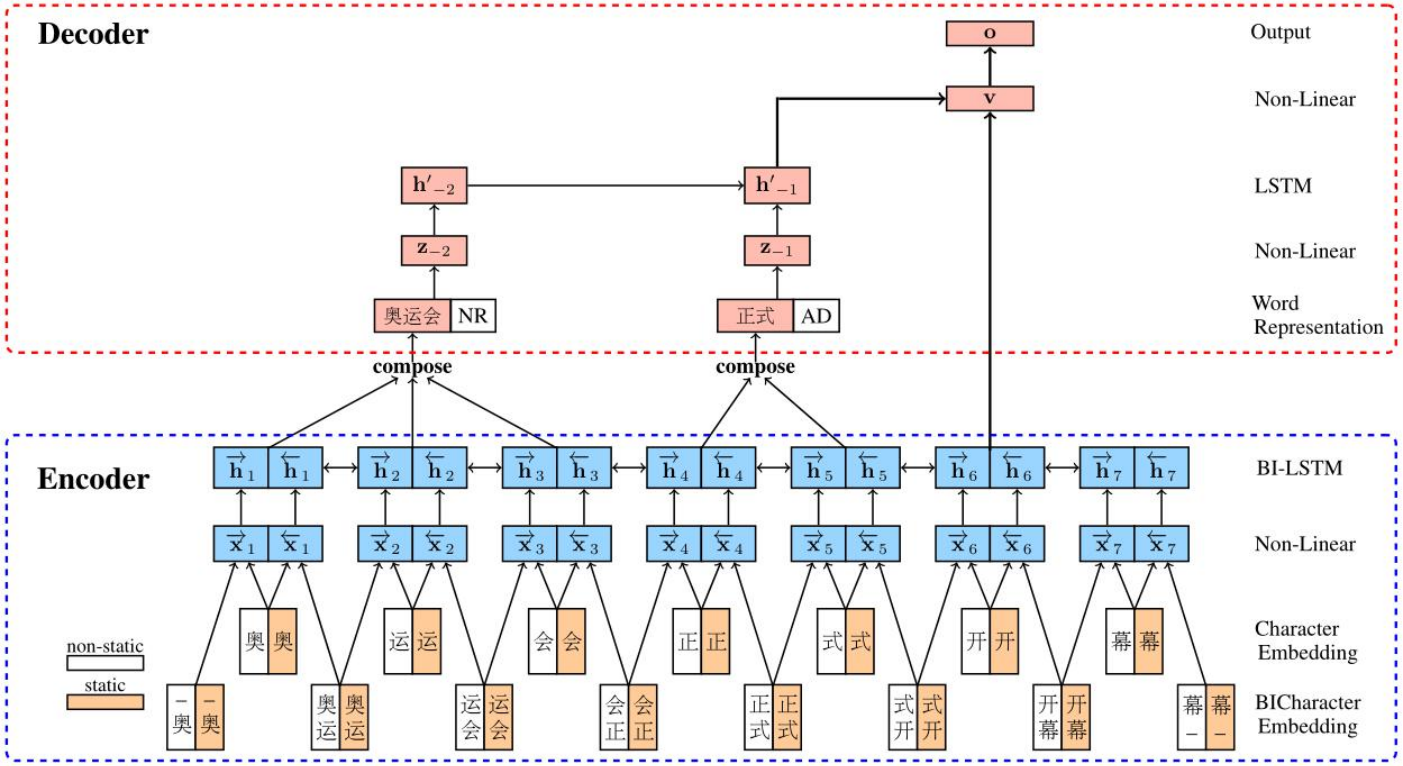
Step	Action	State	
		stack($\cdots w_{-2}\|t_{-2}\ \ w_{-1}\|t_{-1}$)	queue($c_0 c_1 \cdots$)
0	-	ϕ	ao yun ···
1	SEP(NR)	奥(ao)|NR	运(yun) 会(hui) ···
2	APP	奥运(ao yun)|NR	会(hui) 正(zheng) ···
3	APP	奥运会(ao yun hui)|NR	正(zheng) 式(shi) ···
4	SEP(AD)	奥运会(ao yun hui)|NR 正(zheng)|AD	式(shi) 开(kai) 幕(mu)
5	**APP**	**奥运会(ao yun hui)|NR 正式(zheng shi)|AD**	**开(kai) 幕(mu)**
6	SEP(VV)	奥运会(ao yun hui)|NR 正式(zheng shi)|AD 开(kai)|VV	幕(mu)
7	APP	奥运会(ao yun hui)|NR 正式(zheng shi)|AD 开幕(kai mu)|VV	ϕ

Table I

SEP(t)：一个新词的开始，并且标注好这个词的词性
APP：紧接着上一个汉字，与前面的多个汉字组成一个词

例： “奥运会正式开幕”
解码结果： “奥运会|NR 正式|AD 开幕|VV”
动作序列： “SEP(NR) APP APP SEP(AD) APP SEP(VV) APP”

transition-based系统



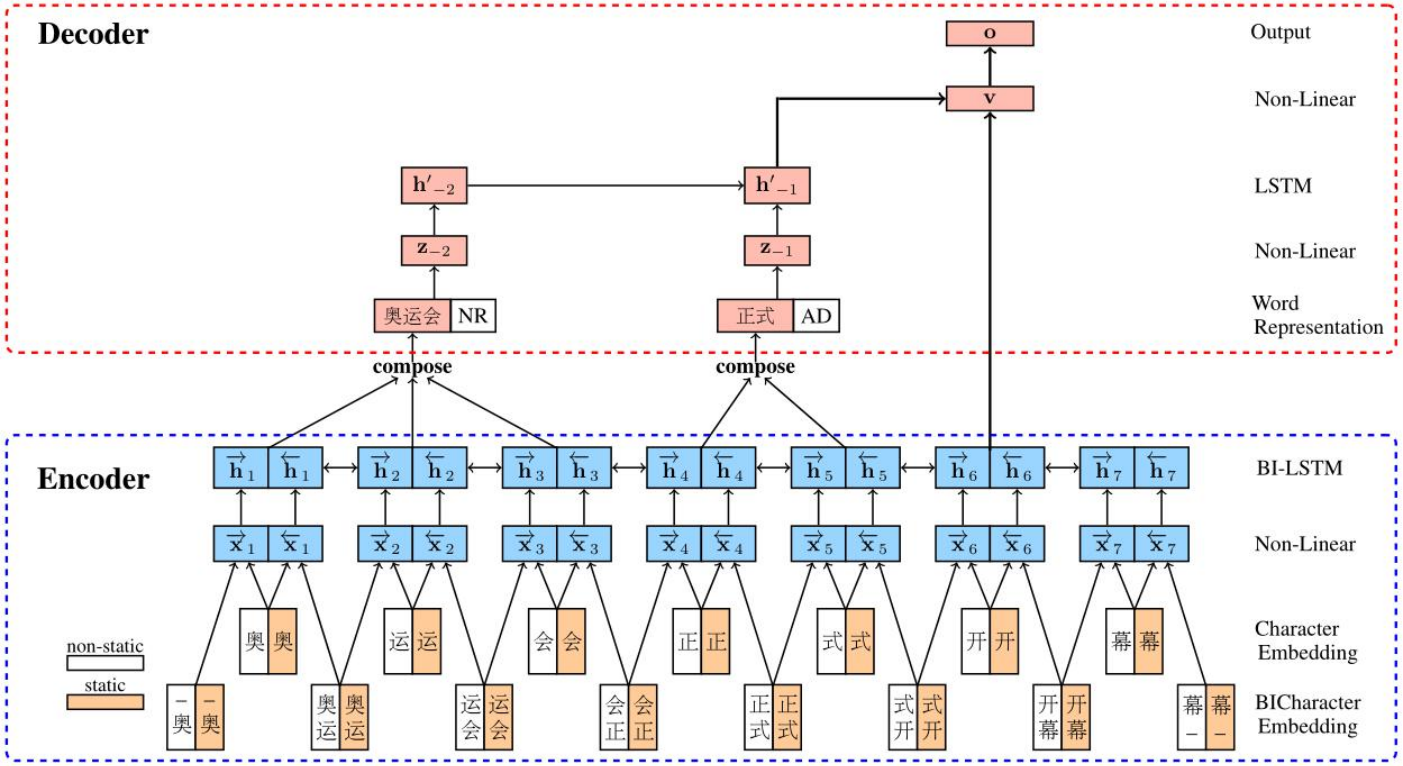
单字和双字的双向编码，为动态编码

采用外部预训练词向量，为静态编码

- Basic Embeddings
- Word-Context Embeddings

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

transition-based系统



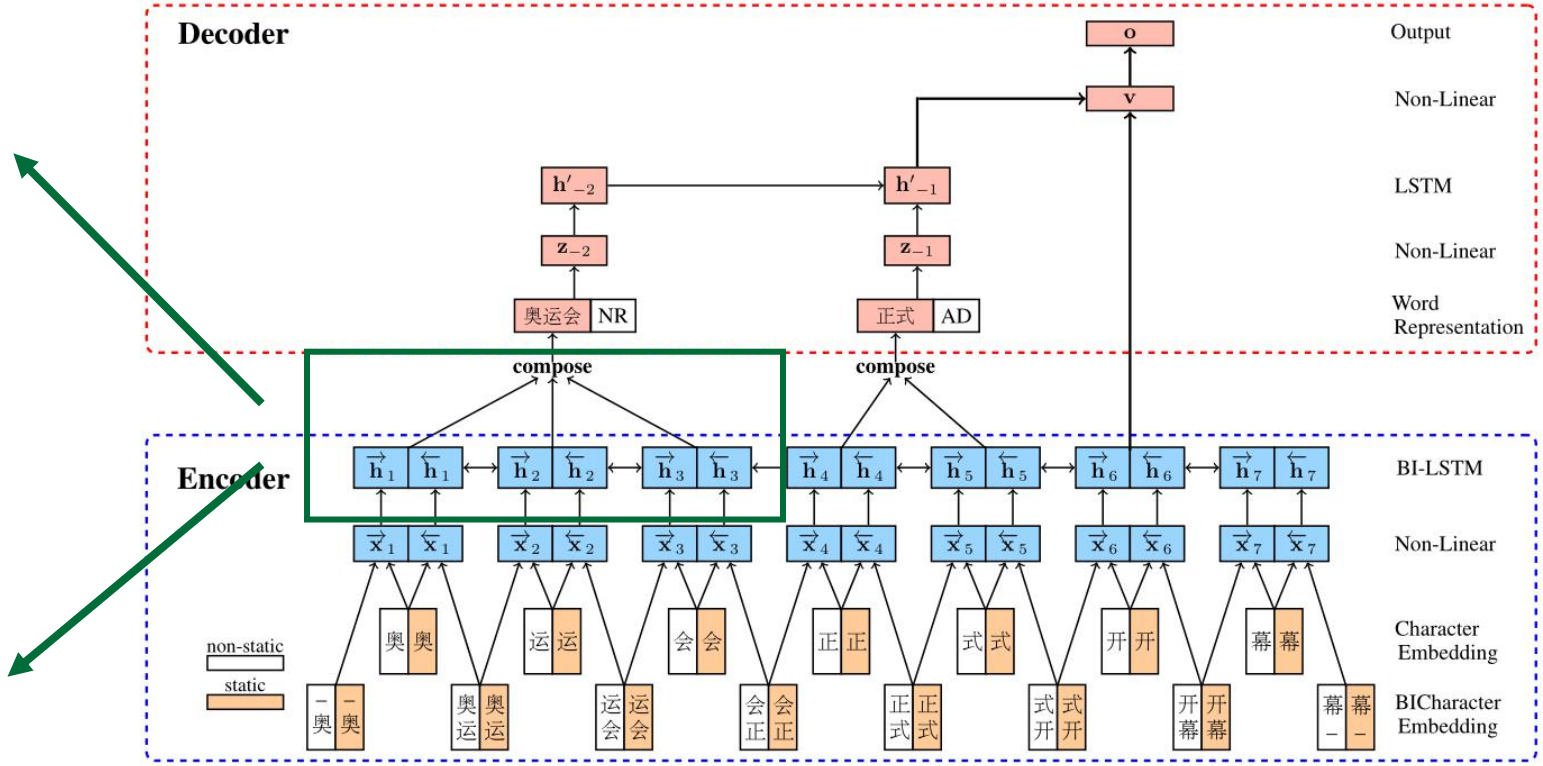
解码端词的表示(Word Representation)

结合解码端LSTM的隐层输出和编码端向量表示进行预测

transition-based系统

典型的Seq2Seq模型需要采用Attention机制，而本文提出的模型不需要采用Attention机制，采用编码端的向量表示。

在分词和词性标注的任务中，词级别的特征异常重要，解码端LSTM是构建在输出的词之上，而不是构建在属于字符级的预测动作序列之上。



[1] Xiaoqing Zheng, Hanyang Chen, Tianyu Xu, “Deep Learning for Chinese Word Segmentation and POS Tagging”, EMNLP 2013: 647-657.

[2] Y . Shao, C. Hardmeier, J. Tiedemann, and J. Nivre, “Character-based joint segmentation and pos tagging for chinese using bidirectional rnn-crf,” arXiv preprint arXiv:1704.01314, 2017.

[3] Meishan Zhang, Nan Y u, and Guohong Fu, “A Simple and Effective Neural Model for Joint Word Segmentation and POS Tagging”, IEEE ACM Trans. Audio Speech Lang. Process. 26(9): 1528-1538 (2018)

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

3

领域前沿

主讲人：李乐凡

当前研究数据

SOTA模型
——WMSEG与TwASP

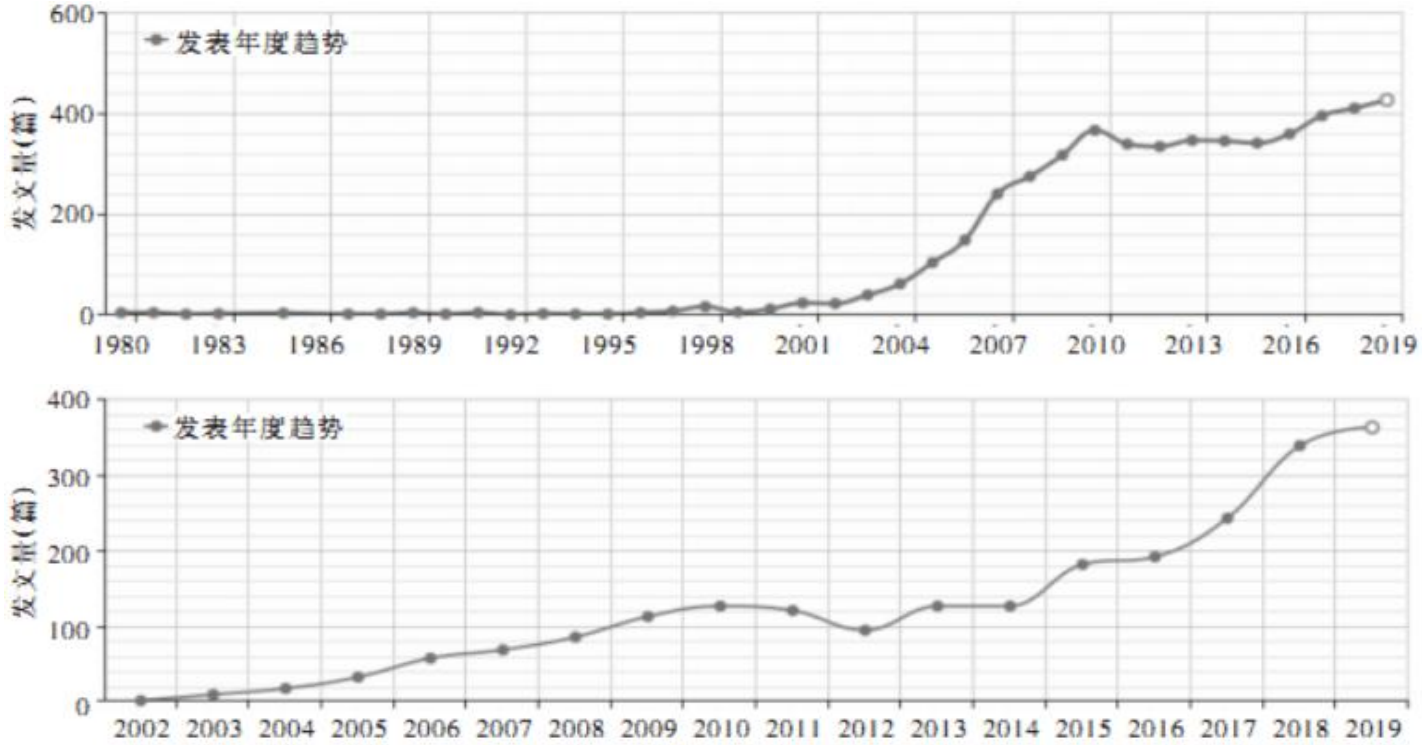
2021最新成果(简介)
——Meta-Seg与GCA-FL

新的瓶颈与发展现状

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

当前研究数据

根据近20年文献资料，中文分词研究自2010年达到小高峰后，热度再次缓步增长。



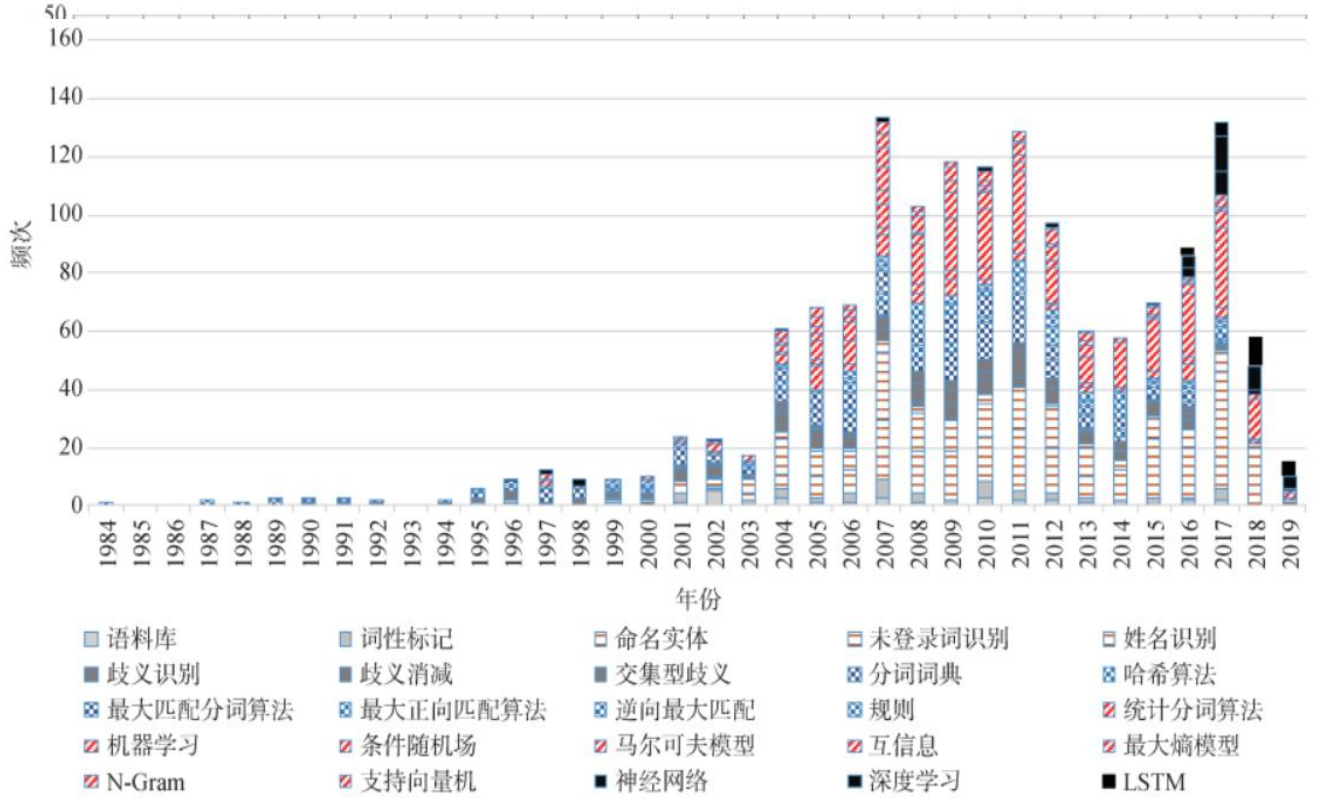
中国知网（上）和Web of Science（下）分词文献数量统计

[1]王佳楠, 梁永全. 中文分词研究综述[J]. 软件导刊, 2021, 20(04):247-252.

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

当前研究数据

所基于的技术变化，统计仅选择总词频高于20的主要技术性关键词的词频分布。



“中文分词文献”部分关键词分布（篇）

- **机械分词算法**

  自1984年至今持续出现在文献中。

- **机器学习算法**

  2004年后被广泛应用，并持续保持较高的关注度。

- **深度学习算法**

  2015年之后相关文献逐渐增多。

但出现的新算法并未替代之前的分词算法。

[2] 唐琳, 郭崇慧, 陈静锋. 中文分词技术研究综述[J]. 数据分析与知识发现, 2020, 4(Z1): 1–17.

SOTA模型

目前中文分词的SOTA（State of the art，最先进）模型，来自创作于我国广州的高科技企业创新工场大湾区人工智能研究院的文章，发表于2020年7月上旬线上举行的第58届自然语言处理领域（NLP）顶级学术会议ACL 2020，目前已开源。

WMSEG：键-值记忆神经网络的中文分词模型

在所有数据集上的表现均超过前人的工作，“把中文分词领域广泛使用的标准数据集上的性能全部刷到了新高。”

TwASP：基于双通道注意力机制的分词及词性标注模型

两模型分别就中文分词和词性标注作出探索，将外部知识（信息）创造性融入分词及词性标注模型，有效剔除分词“噪音”误导，大幅提升处理效果。

执行院长宋彦（作者之一）

SOTA模型——WMSEG：键-值记忆神经网络的中文分词模型

解决OOV（out of vocabulary，未登录词）和歧义两大难题，主要思想是采用键-值记忆神经网络，计算能得出具备更完整语义分词结果的汉字划分方式。特定语境中：

构建词表与分配权重：

歧义消解

部分居民生活水平

未登录词处理

① 据构词能力，找到所有成词组合。如“民”字可能单字成词，作为“居民”的词尾、作为“民生”的词首，或是在“居民生活”的词中成分。

② 利用每个汉字的元组提供的每个字的构词能力，通过加/降权重实现。将找到的全部组合加入到分词模型，进行编码。用非监督方法构建词表，有效利用特定的未标注文本，来提升识别能力。

③ 经神经网络，学习各词对完整表达句意的帮助，从而分配不同权重。最终“部分”、“居民”、“生活”、“水平”被突出，而“分居”、“民生”则被降权。

[3]Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, Yonggang Wang: Improving Chinese Word Segmentation with Wordhood Memory Networks. ACL 2020: 8274-8285

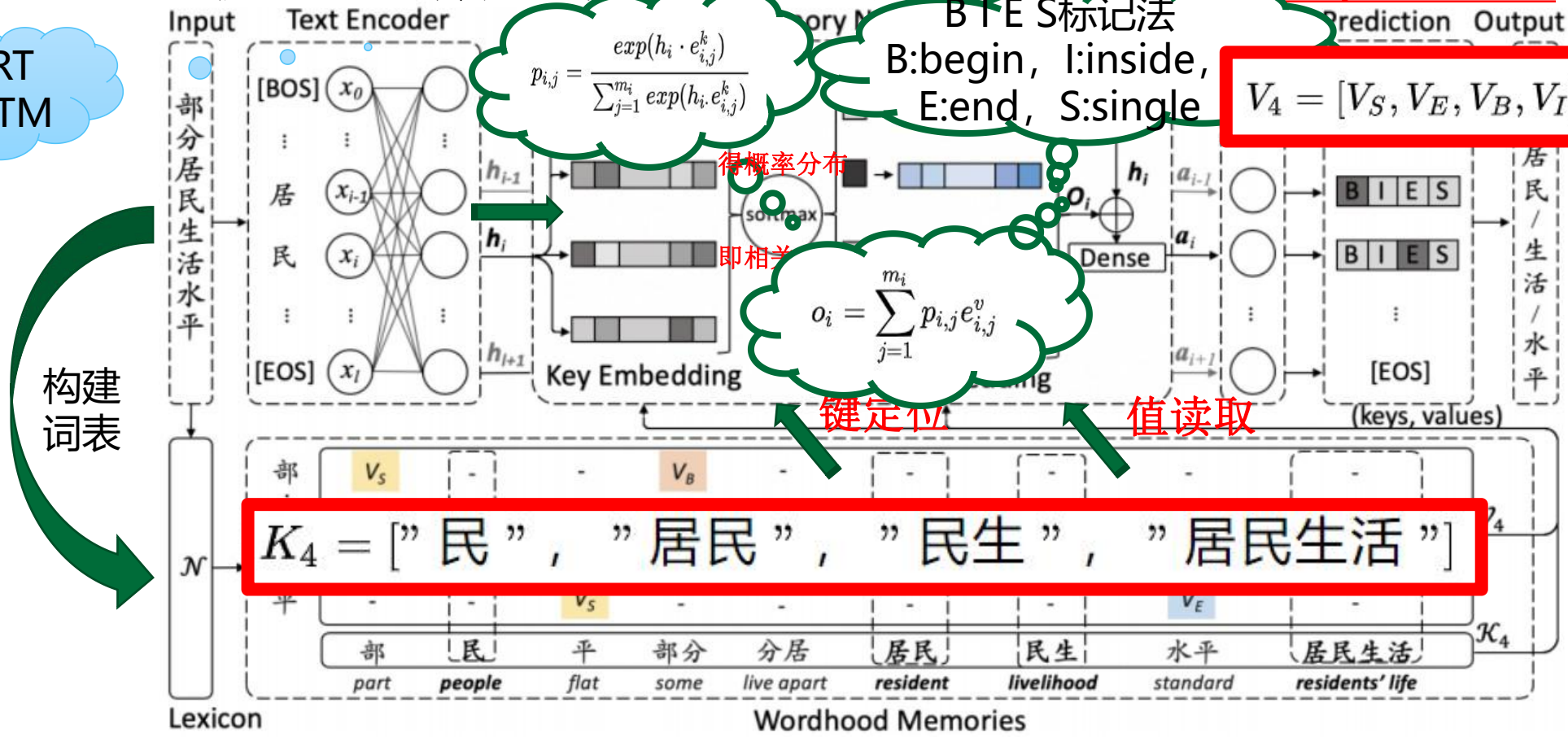
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

SOTA模型——WMSEG：键-值记忆神经网络的中文分词模型

核心思想：传统NER模型的Encoder和Decoder之间加入Memory Networks。

BERT / LSTM

softmax / CRF



模型整体：

$$\hat{y} = \underset{y \in \tau^l}{argmax}\, p(y|X, M(X, N))$$

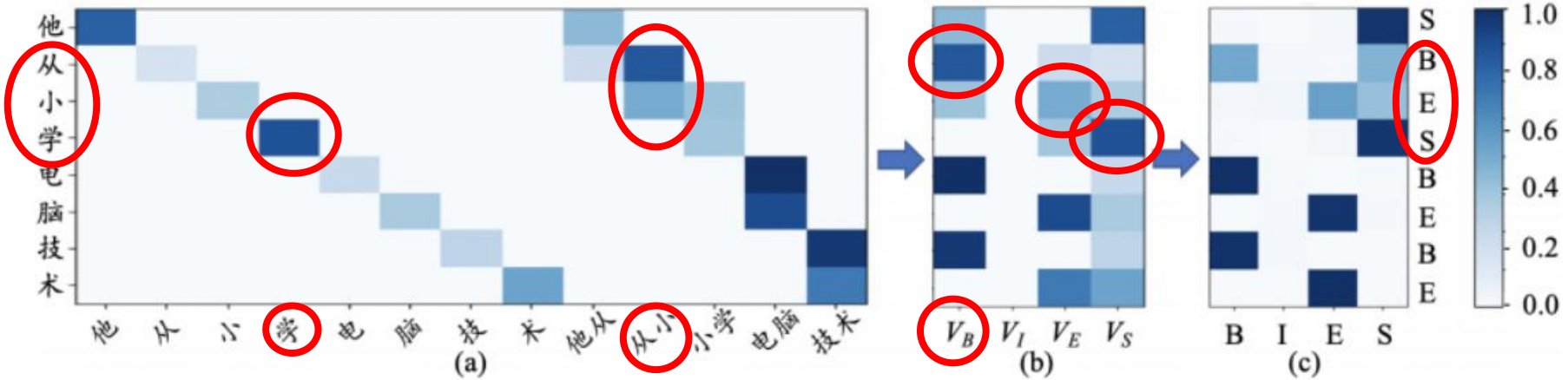
τ：句子所有分词结果的标签集合；
L：句子长度；
ŷ：模型的最好结果；
N：构建的Lexicon；
X：输入的句子；
M：本文模型。

SOTA模型——WMSEG：键-值记忆神经网络的中文分词模型

值读取举例：

他从小学电脑技术

模型对歧义部分“从小学”（有“从/小学”和“从小/学”两种分法）各分法中的n元组“从小”和“学”能够分配更高的权重。



北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

SOTA模型——WMSEG：键-值记忆神经网络的中文分词模型

在主流公开分词模型中加入WM网络进行对比、和前人工作的比较：优化均明显。

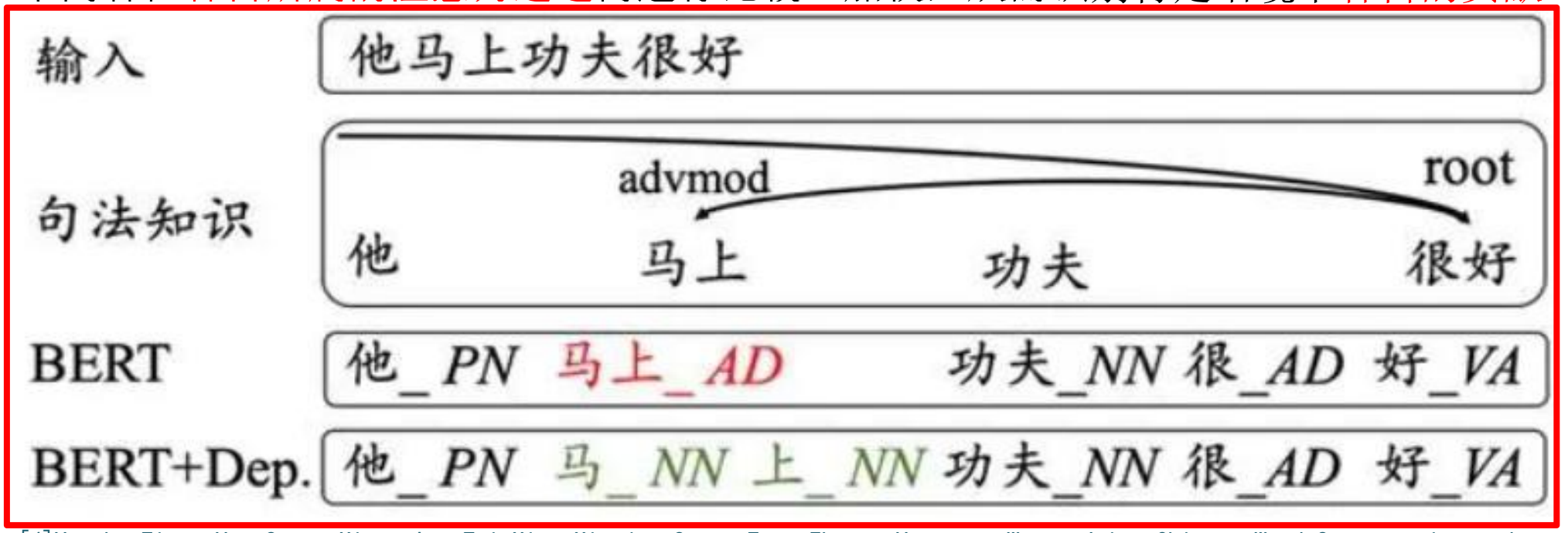
Config		BC		BN		MZ		NW		Web	
En-Dn	WM	F	R_{OOV}	F	R_{OOV}	F	R_{OOV}	F	R_{OOV}	F	R_{OOV}
BL-SM	×	93.73	63.39	93.65	68.88	90.55	**66.95**	93.70	69.57	90.81	55.50
	√	**94.04**	**63.53**	**93.91**	**72.32**	**90.76**	65.65	**93.83**	**72.40**	**91.22**	**56.62**
BL-CRF	×	93.95	65.60	93.87	71.89	90.67	67.13	93.87	72.17	91.12	57.51
	√	**94.21**	**66.81**	**94.11**	**74.22**	**90.95**	**67.29**	**93.96**	**74.38**	**91.49**	**58.37**
BT-SM	×	96.27	80.76	96.88	87.90	94.97	84.45	97.08	89.78	94.82	74.00
	√	**96.41**	**81.15**	**97.00**	**89.47**	**95.10**	**85.48**	**97.24**	**91.96**	**95.00**	**75.51**
BT-CRF	×	96.25	79.04	96.87	89.15	94.94	85.27	96.99	91.34	94.79	75.58
	√	**96.43**	**81.29**	**97.09**	**90.29**	**95.11**	**85.32**	**97.21**	**92.48**	**95.03**	**76.30**
ZEN-SM	×	96.39	79.97	96.95	88.93	95.05	85.14	97.17	91.33	94.03	75.33
	√	**96.45**	**81.34**	**97.03**	**89.78**	**95.06**	**85.60**	**97.21**	**91.73**	**95.08**	**75.60**
ZEN-CRF	×	96.30	80.05	96.97	**90.38**	94.93	85.64	97.10	91.03	94.90	74.98
	√	**96.50**	**80.44**	**97.11**	90.29	**95.13**	**85.96**	**97.24**	**91.68**	**95.04**	**75.74**

	MSR	PKU	AS	CityU	CTB6
Ma et al. (2018)	**98.1**	96.1	96.2	**97.2**	**96.7**
Gong et al. (2019)	97.78	96.15	95.22	96.22	-
Qiu et al. (2019)	98.05	**96.41**	**96.44**	96.91	-
WMSeg (BERT-CRF)	98.28	96.51	96.58	97.80	97.16
WMSeg (ZEN-CRF)	**98.40**	**96.53**	**96.62**	**97.93**	**97.25**

SOTA模型——TwASP：基于双通道注意力机制的分词及词性标注模型

将中文分词和词性标注视作联合任务从而一体化完成。

对自动获取的上下文特征和句法知识，分别加权，预测每个字的分词和词性标签，不同者在各自所属的注意力通道内进行比较、加权，从而识别特定语境下各自的贡献。



[4]Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, Yonggang Wang: Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge. ACL 2020: 8286-8296

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

SOTA模型——TwASP：基于双通道注意力机制的分词及词性标注模型

实验验证：

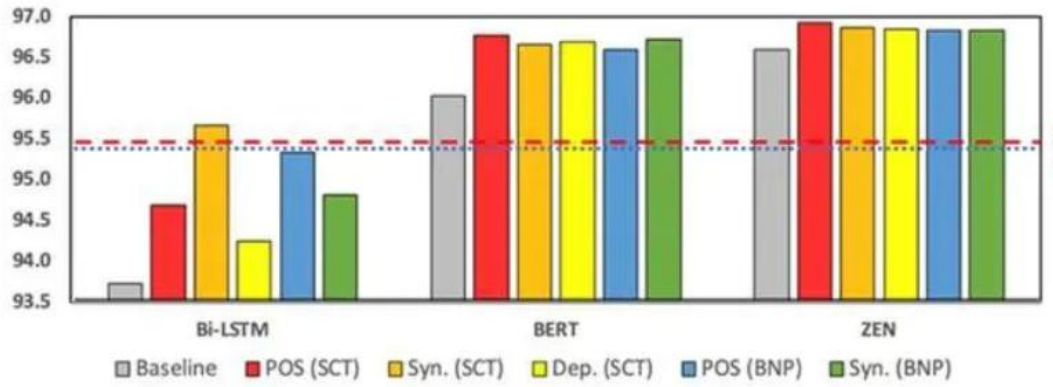
1）模型在所用数据集上均超过了之前的工作：

2）一般领域中，在5个数据集（CTB5，CTB6，CTB7，CTB9，Universal Dependencies）的表现（F值）均超过前人的工作，也大幅度超过斯坦福大学的CoreNLP工具和伯克利大学的句法分析器。

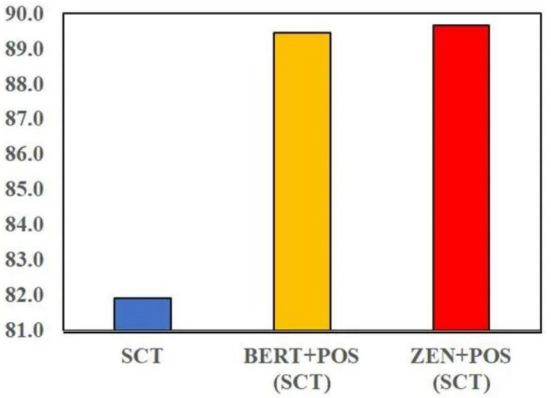
3）跨领域中，模型特别地相对于斯坦福大学的CoreNLP工具有近10个百分点提升。

	CTB5	CTB6	CTB7	CTB9	UD1	UD2
Kurita et al. (2017)	94.84	-	91.25	-	-	-
Shao et al. (2017)	94.38	-	-	**92.34**	**89.75**	**89.42**
Zhang et al. (2018)	94.95	92.51	91.87	-	-	-
斯坦福CoreNLP工具	95.49	90.85	92.73	88.23	0.00	36.11
伯克利句法分析器	**95.50**	**94.43**	**92.95**	88.09	0.00	27.16
BERT+POS (SCT)	96.77	94.82	94.12	94.87	95.60	95.46
ZEN + POS (SCT)	**96.92**	**94.87**	**94.20**	**94.88**	**95.69**	**95.49**

典型测试表现



一般领域的先进性（最常见的CTB5上的结果）



跨领域的先进性

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

2021最新成果——截至2021年10月12日

DBLP 2021			
机械方法	1	W-core Transformer Model for Chinese Word Segmentation	变换器（Transformer）模型、窗核（W-core）
	2	Corpus Annotation System Based on HanLP Chinese Word Segmentation	弹性搜索
	3	More than Text: Multi-modal Chinese Word Segmentation	多模态、变换器（Transformer）模型
	4	Span Labeling Approach for Vietnamese and Chinese Word Segmentation	跨度标记方法
		Pre-training with Meta Learning for Chinese Word Segmentation	元学习
		Federated Chinese Word Segmentation with Global Character Associations	联邦学习、深度学习
深度学习	3	Bidirectional LSTM-CRF Attention-based Model for Chinese Word Segmentation	注意机制、双向长短期记忆+条件随机场（Bi-LSTM-CRF）
	4	Research on Chinese Word Segmentation Based on Conditional Random Fields	条件随机场（CRF）、域自适应、域分割、逆向最大匹配
	5	Exploring Word Segmentation and Medical Concept Recognition for Chinese Medical Texts	长短期记忆（BiLSTM）、变换器双向编码表示（BERT）、中文预训练语言模型ZEN
	6	Enhancing Chinese Word Segmentation via Pseudo Labels for Practicability	半监督、伪标签、神经网络
	7	Hybrid Feature Fusion Learning Towards Chinese Chemical Literature Word Segmentation	混合特征融合、知识提取

ACL 2021
2篇

2021最新成果——Meta-Seg：基于元学习的中文分词预训练模型

Models	PKU	MSRA	CITYU	AS	CKIP	NCC	SXU	CTB6	CNC	Avg.
Chen et al. (2017)	94.32	96.04	95.55	94.64	94.26	92.83	96.04	-	-	-
Ma et al. (2018)	96.10	97.40	97.20	96.20	-	-	-	96.70	-	-
He et al. (2019)	95.78	97.35	95.60	95.47	95.73	94.34	96.49	-	-	-
Gong et al. (2019)	96.15	97.78	96.22	95.22	94.99	94.12	97.25	-	-	-
Yang et al. (2019)	95.80	97.80	-	-	-	-	-	96.10	-	-
Meng et al. (2019)	96.70	98.30	97.90	96.70	-	-	-	-	-	-
Yang (2019)	96.50	98.40	-	-	-	-	-	-	-	-
Duan and Zhao (2020)	95.50	97.70	96.40	95.70	-	-	-	-	-	-
Huang et al. (2020)	**97.30**	98.50	97.80	97.00	-	-	97.50	97.80	97.30	-
Qiu et al. (2020)	96.41	98.05	96.91	96.44	96.51	96.04	97.61	-	-	-
Tian et al. (2020)	96.53	98.40	97.93	96.62	-	-	-	97.25	-	-
BERT-Base (ours)	96.72	98.25	98.19	96.93	96.49	96.13	97.61	97.85	97.45	97.29
METASEG (w/o fine-tune)	96.76	98.02	98.12	**97.04**	**96.81**	97.21	97.51	97.87	97.25	97.40
METASEG	96.92	**98.50**	**98.20**	97.01	96.72	**97.24**	**97.88**	**97.89**	**97.55**	**97.55**

[5]Zhen Ke, Liang Shi, Songtao Sun, Erli Meng, Bin Wang, Xipeng Qiu: Pre-training with Meta Learning for Chinese Word Segmentation. NAACL-HLT 2021: 5514-5523

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

2021最新成果——GCA-FL：基于全局字符关联机制联邦学习的中文分词

用于数据隔离的场景下提升模型在中文分词性能。

模型：存于服务器端

数据：存于节点，节点间孤立不可见。

① 采用联邦学习(federated learning，FL)进行分布式学习，在保证隐私安全与合法的前提下，解决数据孤立的问题，实现共同建模。

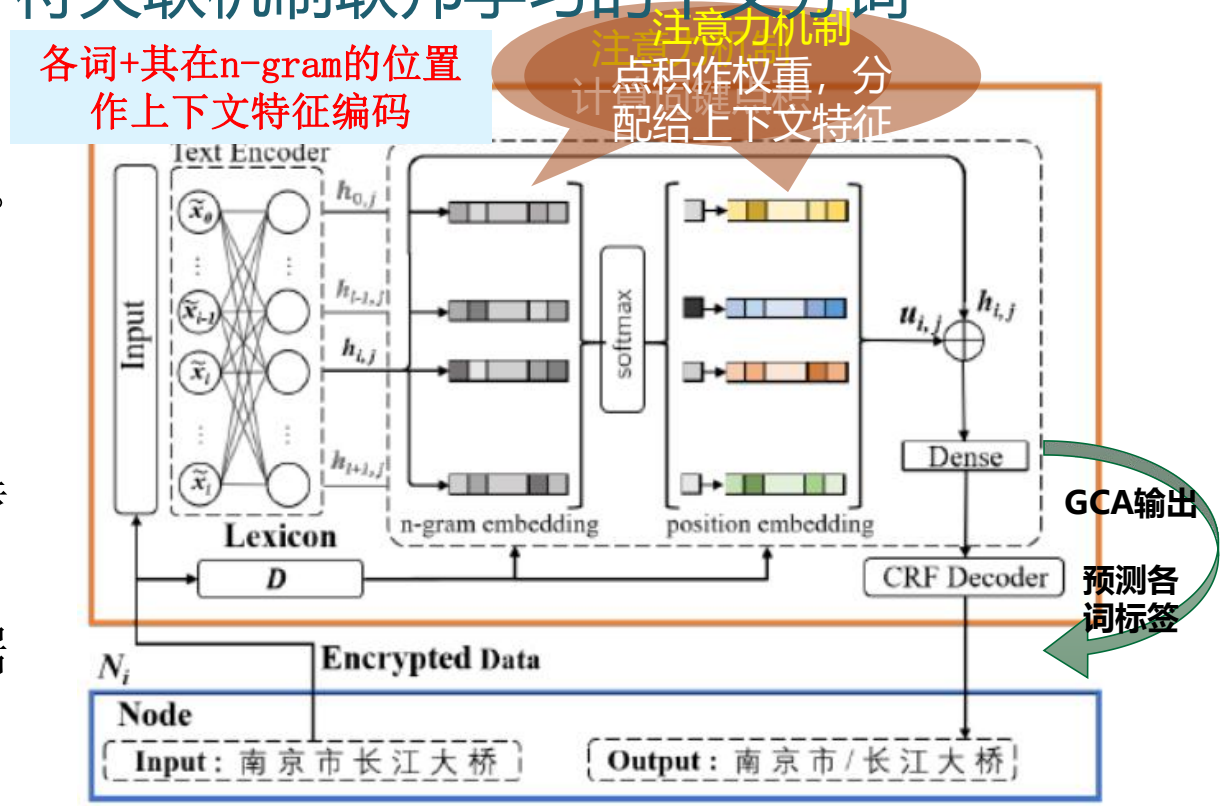
② 使用一种全局字符关联机制(Global Character Associations，GCA)的方法，增强模型对数据孤立情景的中文分词任务处理高性能。

模型和各节点通信过程：

① 节点——加密数据→服务器。

② 服务器端模型据此前向计算，并传输解码后的分词标签给节点。

③ 节点据此计算损失，最后模型根据损失反向传播计算梯度并更新参数。



联邦学习的训练过程

服务器端模型结构

[6]Yuanhe Tian, Guimin Chen, Han Qin, Yan Song: Federated Chinese Word Segmentation with Global Character Associations. ACL/IJCNLP (Findings) 2021: 4306-4313

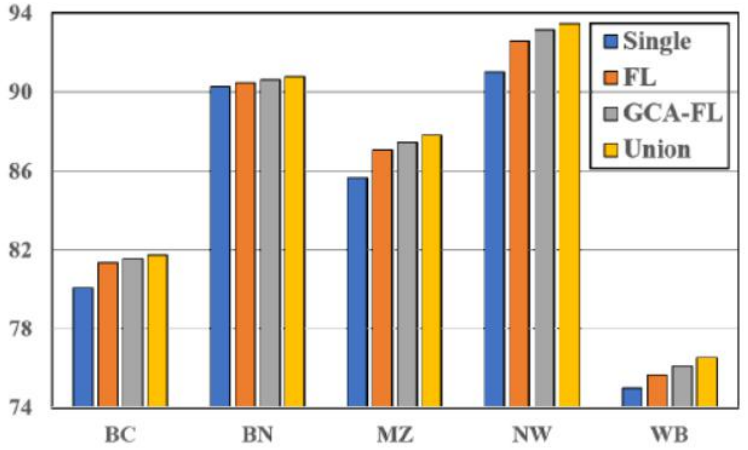
2021最新成果——GCA-FL：基于全局字符关联机制联邦学习的中文分词

	BC	BN	MZ	NW	WB	Avg.
Single	97.13	96.97	96.21	97.84	94.83	96.60
Union	97.80	97.49	96.74	98.44	95.30	97.15
FL	97.49	97.22	96.54	98.15	95.03	96.89
GCA-FL	**97.76**	**97.40**	**96.74**	**98.43**	**95.29**	**97.12**

(a) BERT

	BC	BN	MZ	NW	WB	Avg.
Single	97.43	97.38	96.33	98.11	95.14	96.88
Union	97.88	97.79	97.23	98.61	96.38	97.58
FL	97.62	97.56	96.88	98.44	95.88	97.28
GCA-FL	**97.83**	**97.76**	**97.01**	**98.50**	**95.94**	**97.41**

(b) ZEN 2.0



实验结果：

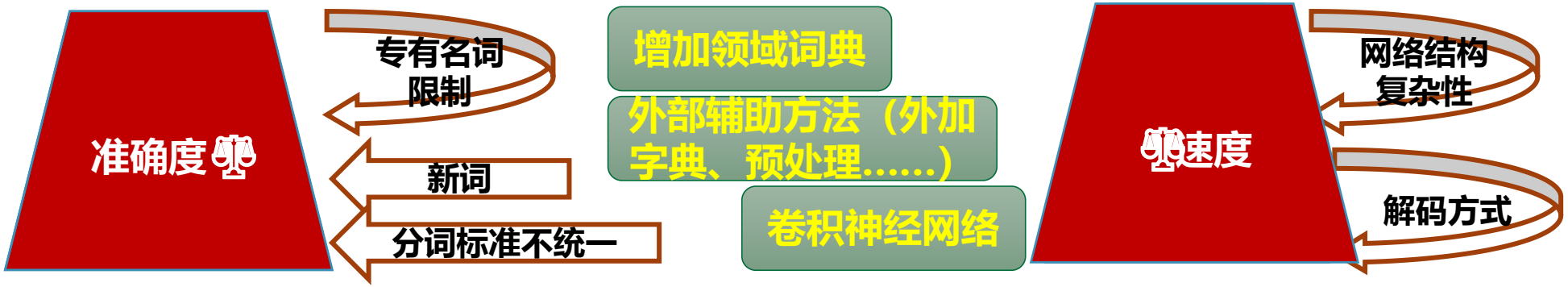
1）实验结果表明了该方法的高效，优于大部分不同的基础模型，其中包括一些设计良好的联邦学习框架。下表是五个基准数据集上的模型性能。

2）此外，下图表现了模型在五个基准数据集上未登录词的召回率，通过分析模型在OOV的问题解决表现验证了联邦学习和全局字关联机制的有效性。

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

新的瓶颈与发展现状

- 中文分词新的瓶颈



- 现状与展望　　日趋成熟
  - **基于词典分词的机械分词方法：**简捷，存在领域局限与歧义，局外词汇识别差。
  - **基于统计分词的监督学习算法：**转换分词为序列标注，改进歧义等问题，CRF和HMM模型成为统计分词的主要方法。
  - 近年，神经网络的出现使分词准确度有了极大提高，但在Bi-LSTM+CRF算法应用于分词领域后，准确度的提升空间逐步变小。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

4

技术平台及应用场景

主讲人：蒋凌昀

pkuseg

- 多领域分词
- 高分词准确率
- 支持用户自训练模型
- 支持词性标注

pkuseg

细领域分词

```
import pkuseg

seg = pkuseg.pkuseg(model_name='medicine')
text = seg.cut('我爱北京天安门')
print(text)
```

词性标注

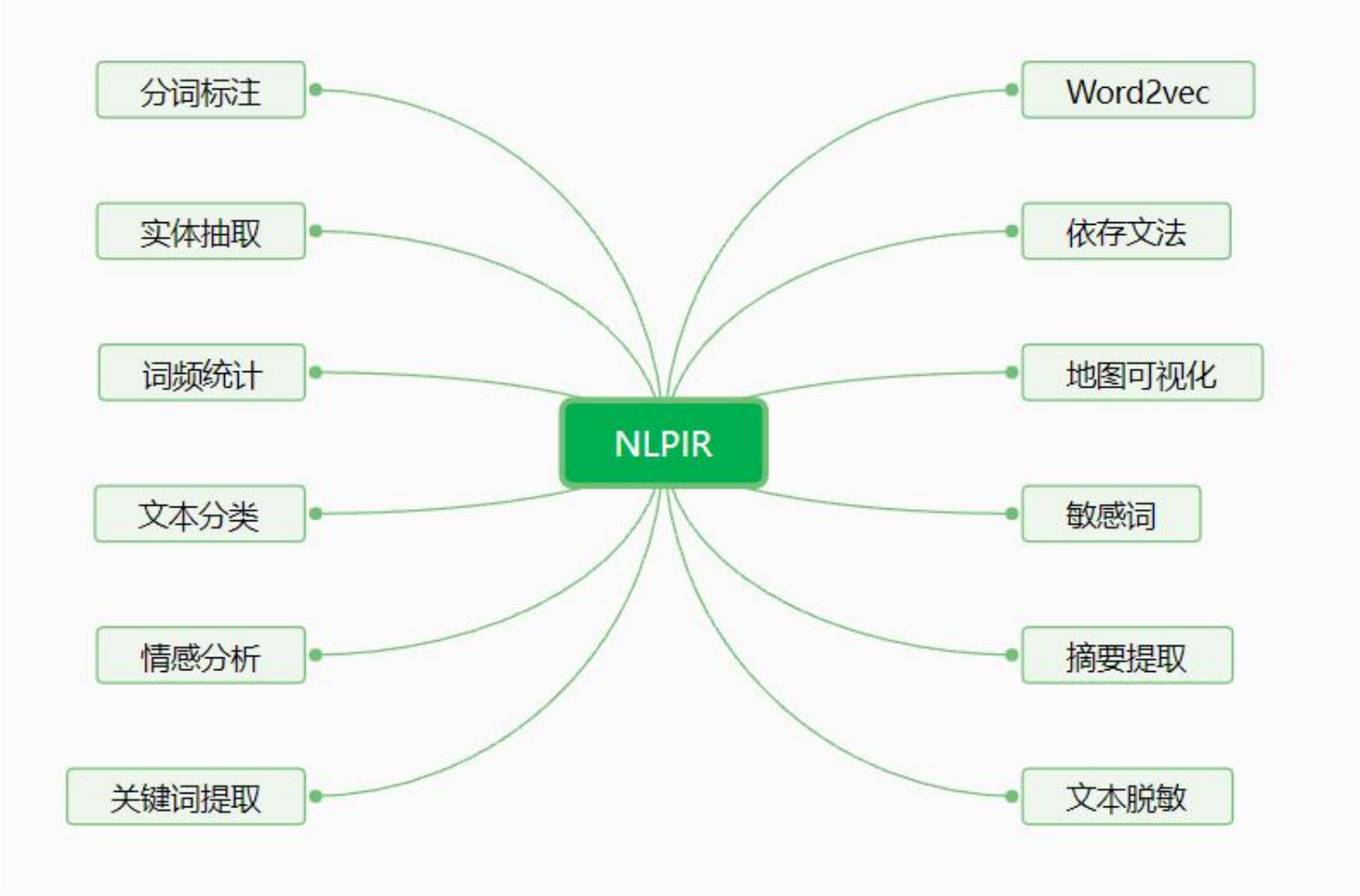
```
import pkuseg

seg = pkuseg.pkuseg(postag=True)
text = seg.cut('我爱北京天安门')
print(text)
```

自训练模型

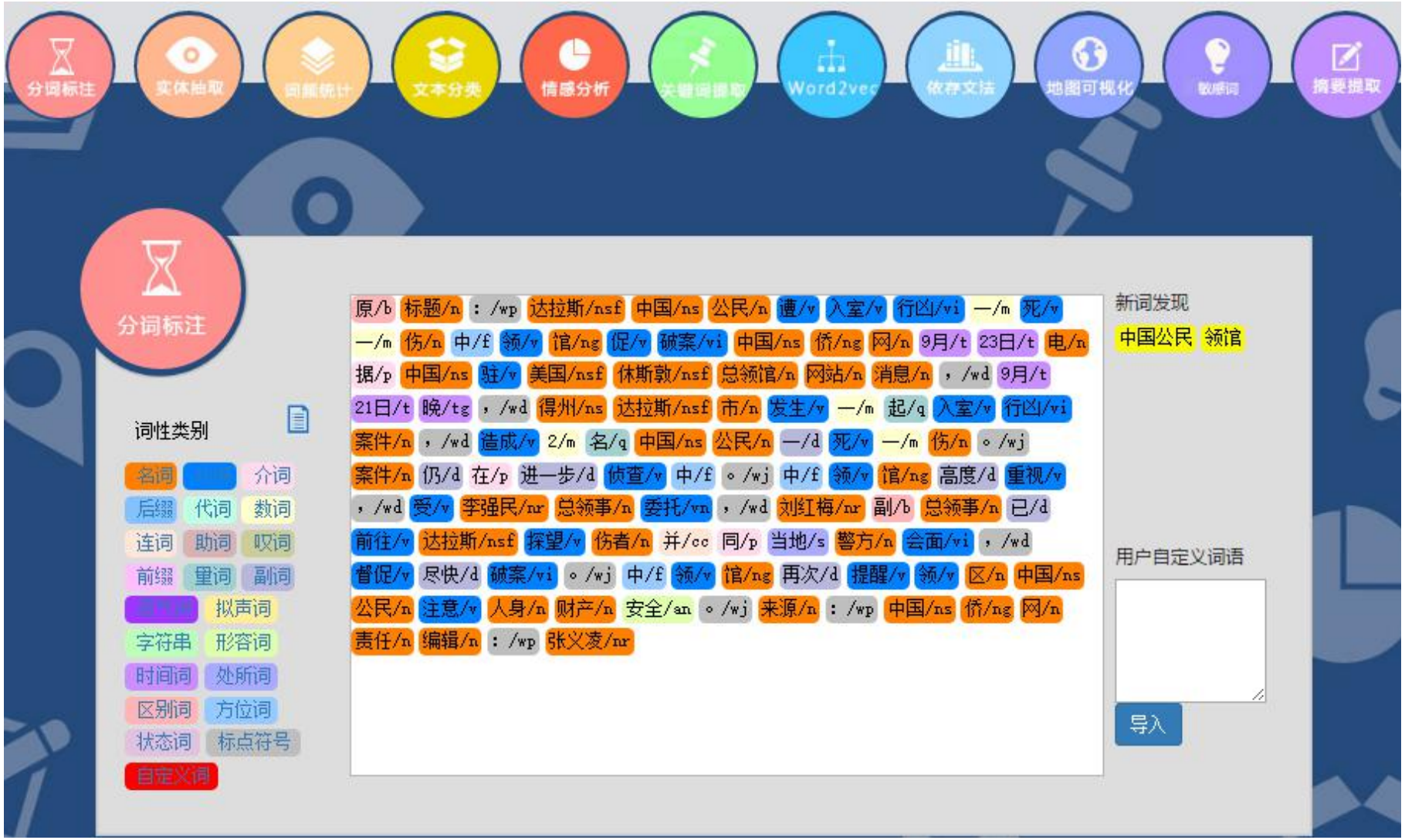
```
pkuseg.train(trainFile, testFile, savedir, train_iter = 20, init_model = None)
```

NLPIR

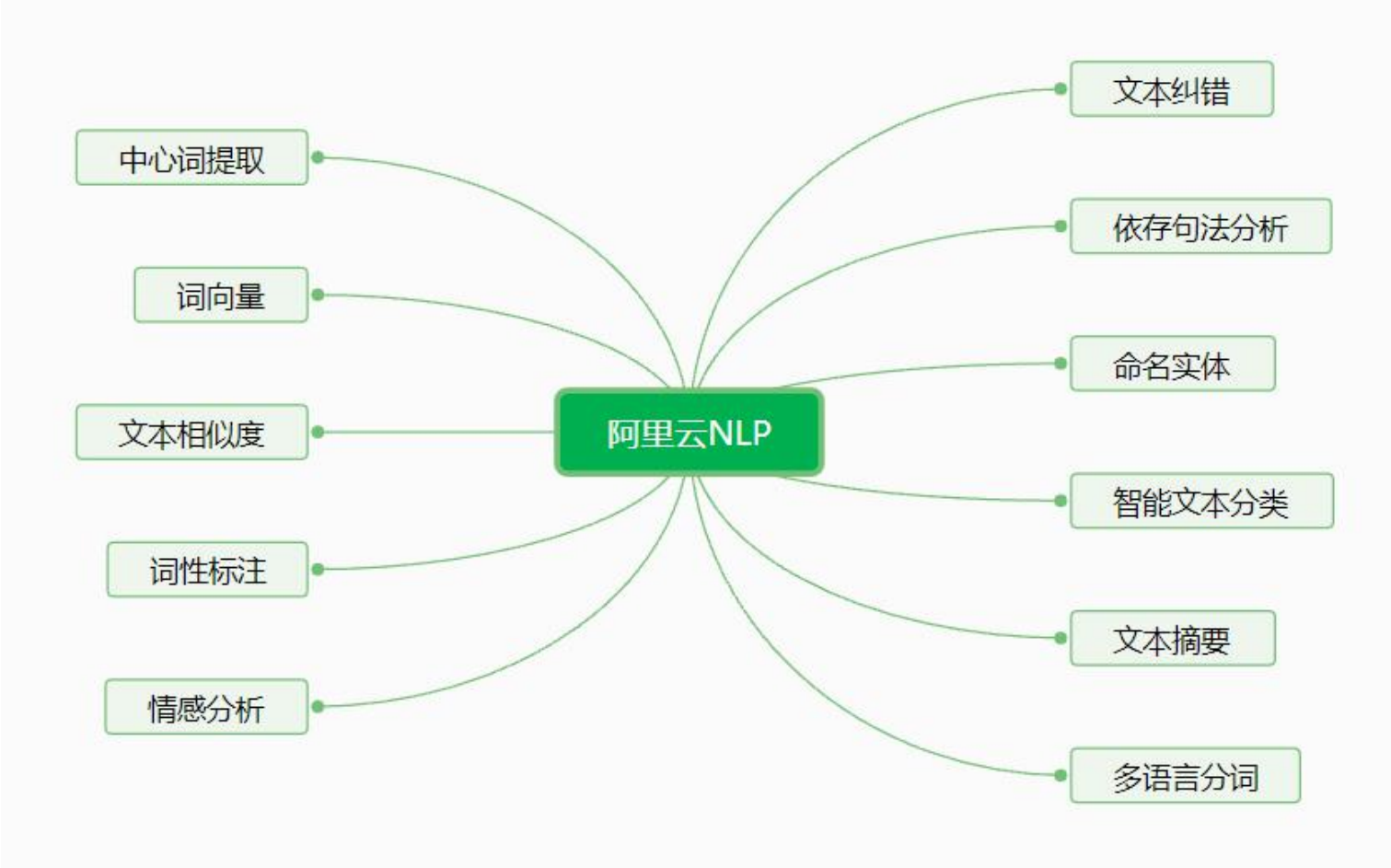

分词标注
实体抽取
词频统计
文本分类
情感分析
关键词提取
NLPIR
Word2vec
依存文法
地图可视化
敏感词
摘要提取
文本脱敏

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

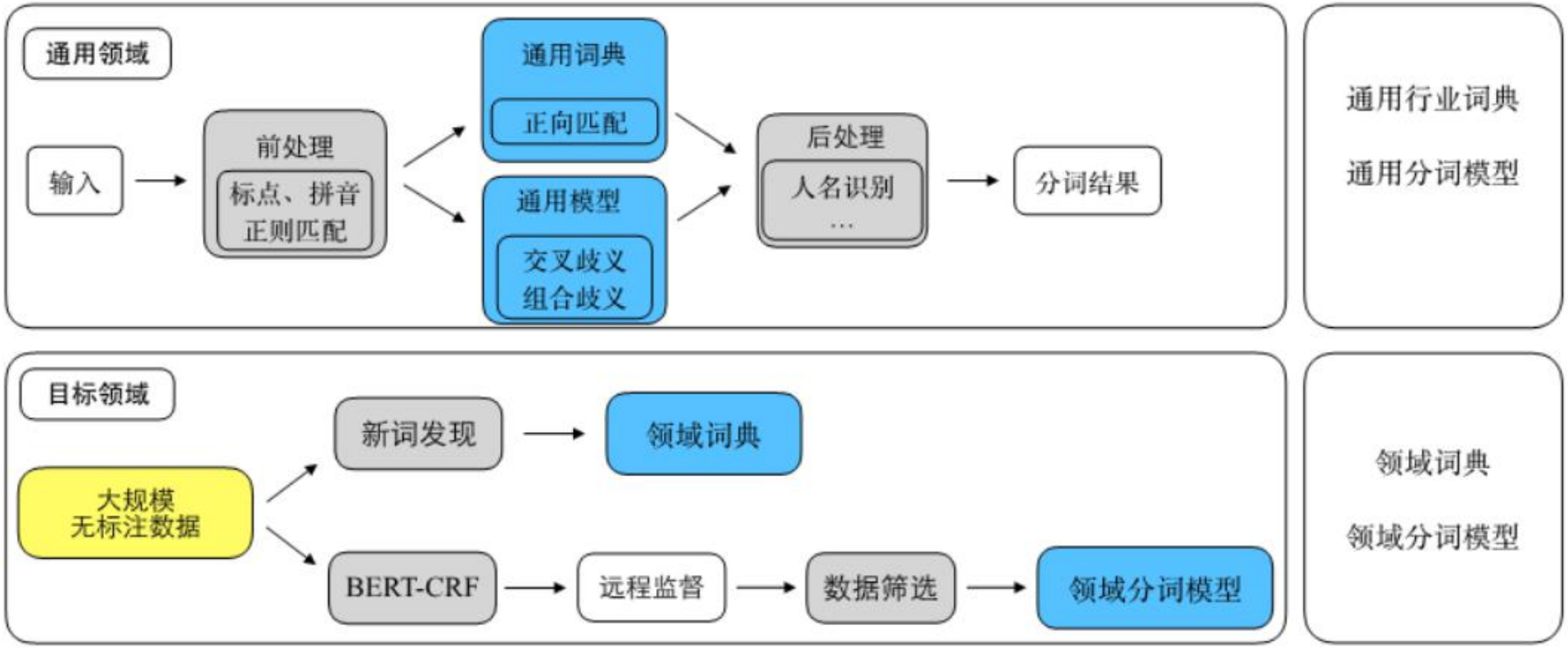
NLPIR



阿里云NLP


阿里云NLP
中心词提取
词向量
文本相似度
词性标注
情感分析
文本纠错
依存句法分析
命名实体
智能文本分类
文本摘要
多语言分词

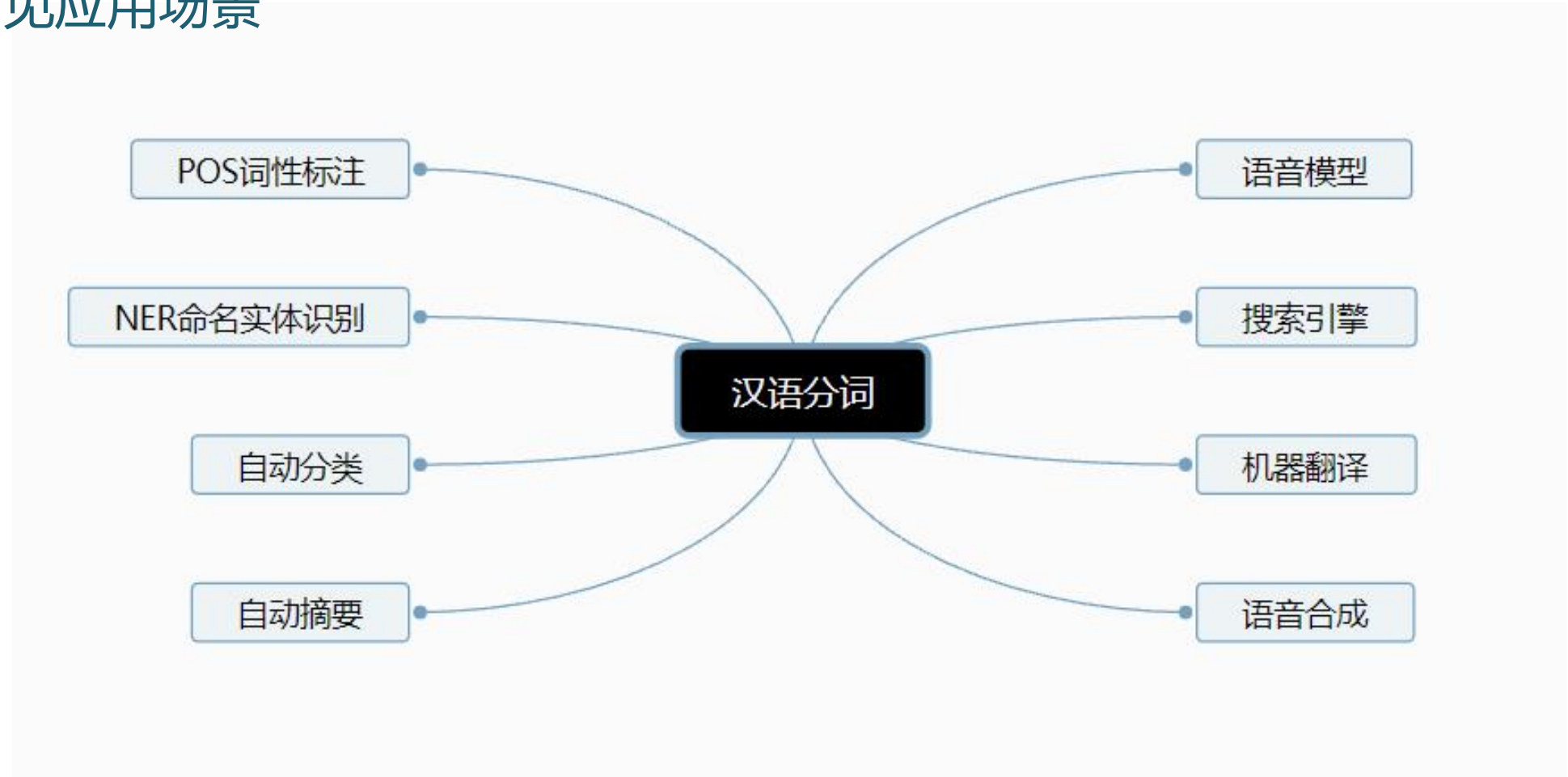
阿里云NLP


通用领域
输入
前处理
标点、拼音
正则匹配
通用词典
正向匹配
通用模型
交叉歧义
组合歧义
后处理
人名识别
…
分词结果
通用行业词典
通用分词模型
目标领域
大规模
无标注数据
新词发现
领域词典
BERT-CRF
远程监督
数据筛选
领域分词模型
领域词典
领域分词模型

阿里云NLP

Query	开源IK分词	OpenSearch电商分词
苹果六s手机壳剥离	苹果 六 手 机壳 玻璃	苹果 六 s 手机 壳 玻璃
冬季素颜霜	冬季 素 颜 霜	冬季 素颜 霜
抖音玩具大人爱心钻石	抖 音 玩具 大人 爱心 钻石	抖音 玩具 大人 爱心 钻石
百雀羚套装旗舰店	百 雀 羚 套装 旗舰 店	百雀羚 套装 旗舰店
925银耳饰	925 银耳 饰	925 银 耳饰
火锅九块九包邮	火锅 九 块 九 包 邮	火锅 九块九 包邮

常见应用场景



北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

翻译技术

Baidu 翻译

Google Translate

有道翻译

嗨 hi 腾讯翻译君

语音助手

汉语分词与标注

- 提高生产力
- 技术成熟
- 复杂性

未来展望

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5

Demo展示

—— 基于中文分词对比分析网络新闻标题

主讲人：刘家昌

基于中文分词对比分析网络新闻标题

数据集、分词工具


```
[{
    "title": "岸田文雄当选自民党新任总裁，并将出任第100任日本首相，中方回应",
    "cate": "china",
    "date": "2021-09-29 15:32:12",
    "keywords": "岸田文雄",
    "brief": "29日，日本前外务大臣、自民党前政调会长岸田文雄当选自民党新任总裁。新总裁任期为3年，至2024年9月。10月4日，岸田文雄将在临时国会上正式出任第100任日本首相，并组建新内阁。",
    "url": "https://news.cctv.com/2021/09/29/ARTIbOJTjuLH6XxPtOxCcWGi210929.shtml"
  },
  ...]
```


央视网（国内、国际）	1500
网易新闻（社会，国际）	347
头条新闻（热点）	1674

Default	MSRA	CTB8	PKU	WEIBO	All Average
jieba	81.45	79.58	81.83	83.56	81.61
THULAC	85.55	87.84	92.29	86.65	88.08
pkuseg	87.29	91.77	92.68	93.43	91.29

https://github.com/lancopku/pkuseg-python

基于中文分词对比分析网络新闻标题

分词结果统计分析

点出关键信息：

名词、动词、数词、地名、简称、人名

平均高频词分布方差	
央视网	1.29E-05
网易新闻	3.29E-05
头条新闻	5.82E-05

的	助词	179	0.1193	的	助词	80	0.231	的	助词	808	0.483
一	数词	120	0.0800	中国	地名	73	0.210	个	量词	239	0.143
中国	地名	114	0.0760	美	简称	41	0.118	是	动词	237	0.142
在	介词	103	0.0687	美国	地名	36	0.104	不	副词	214	0.128
人	名词	89	0.0593	被	介词	31	0.089	了	助词	199	0.119
美国	地名	88	0.0587	不	副词	29	0.084	有	动词	195	0.116
被	介词	83	0.0553	一	数词	28	0.081	一	数词	190	0.114
将	副词	72	0.0480	是	动词	28	0.081	岁	量词	176	0.105
国家	名词	72	0.0480	在	介词	25	0.072	被	介词	173	0.103
已	副词	72	0.0480	台湾	地名	25	0.072	了	语气词	169	0.101
例	量词	72	0.0480	了	语气词	24	0.069	后	方位词	158	0.094
病例	名词	65	0.0433	回应	动词	23	0.066	人	名词	153	0.091
新增	动词	63	0.0420	有	动词	22	0.063	为何	代词	144	0.086
和	连词	57	0.0380	阿富汗	地名	21	0.061	你	代词	128	0.076
不	副词	57	0.0380	了	助词	21	0.061	他	代词	108	0.065
为	动词	57	0.0380	大陆	名词	18	0.052	在	介词	108	0.065
确诊	动词	57	0.0380	人	名词	16	0.046	主席	名词	103	0.062
北京	地名	52	0.0347	拜登	动词	16	0.046	什么	代词	99	0.059
美	简称	52	0.0347	名	量词	15	0.043	年	量词	98	0.059
新冠	名词	51	0.0340	台	简称	15	0.043	毛	人名	93	0.056

基于中文分词对比分析网络新闻标题

分词结果统计分析

PKUSEG
vs
NLPIR.ICTCLAS

的	助词	179	0.1193	一	m	131	0.0873
一	数词	120	0.0800	***新***	***a***	***116***	***0.0773***
中国	地名	114	0.0760	在	p	112	0.0747
在	介词	103	0.0687	人	n	109	0.0727
人	名词	89	0.0593	中国	ns	106	0.0707
美国	地名	88	0.0587	例	q	83	0.0553
被	介词	83	0.0553	美国	ns	79	0.0527
将	副词	72	0.0480	将	d	76	0.0507
国家	名词	72	0.0480	不	d	65	0.0433
已	副词	72	0.0480	病例	n	65	0.0433
例	量词	72	0.0480	已	d	63	0.0420
病例	名词	65	0.0433	***冠***	***n***	***62***	***0.0413***
新增	动词	63	0.0420	国家	n	60	0.0400
和	连词	57	0.0380	确诊	v	59	0.0393
不	副词	57	0.0380	美	b	58	0.0387
为	动词	57	0.0380	和	c	55	0.0367
确诊	动词	57	0.0380	新增	v	54	0.0360
北京	地名	52	0.0347	名	q	53	0.0353
美	简称	52	0.0347	发布	v	48	0.0320
新冠	名词	51	0.0340	对	p	47	0.0313

的	助词	808	0.483	不	d	266	0.1589
个	量词	239	0.143	一	m	251	0.1499
是	动词	237	0.142	个	q	246	0.1470
不	副词	214	0.128	人	n	229	0.1368
了	助词	199	0.119	后	f	209	0.1249
有	动词	195	0.116	***么***	***nr***	***207***	***0.1237***
一	数词	190	0.114	年	q	189	0.1129
岁	量词	176	0.105	了	y	188	0.1123
被	介词	173	0.103	岁	q	186	0.1111
了	语气词	169	0.101	了	u	183	0.1093
后	方位词	158	0.094	大	a	144	0.0860
人	名词	153	0.091	为何	r	144	0.0860
为何	代词	144	0.086	***什***	***n***	***144***	***0.0860***
你	代词	128	0.076	你	r	128	0.0765
他	代词	108	0.065	在	p	125	0.0747
在	介词	108	0.065	这	r	113	0.0675
主席	名词	103	0.062	他	r	108	0.0645
什么	代词	99	0.059	主席	n	103	0.0615
年	量词	98	0.059	毛	nr	95	0.0568
毛	人名	93	0.056	最	d	93	0.0556

基于中文分词对比分析网络新闻标题

分词结果统计分析

点出关键信息：

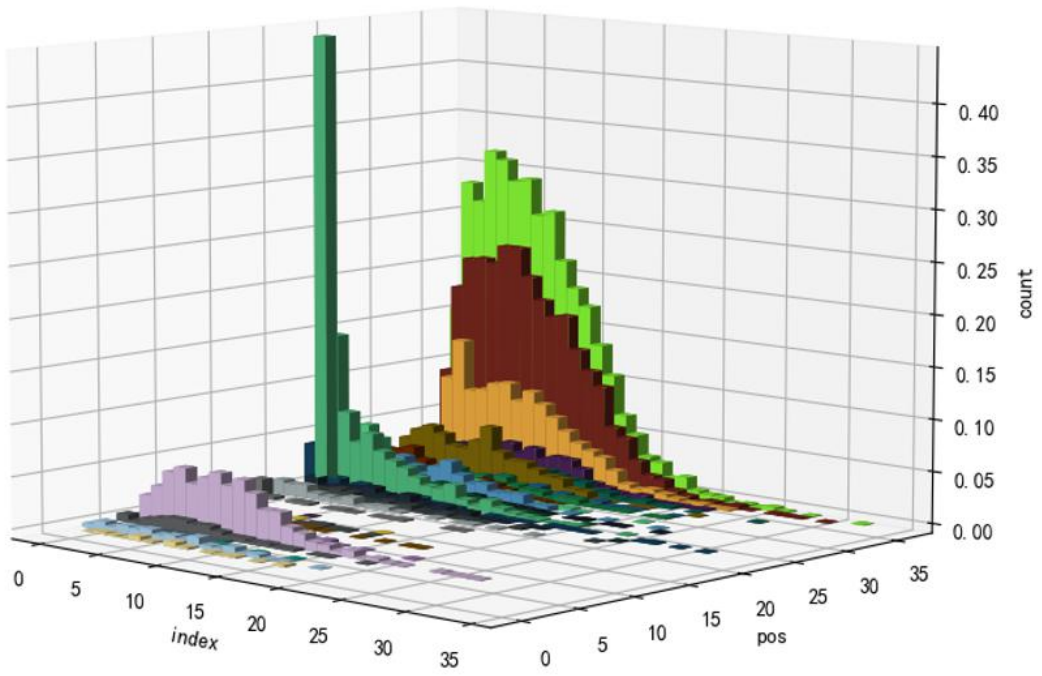
名词、动词、数词、地名、简称、人名

n,名词	5303	3.535	n,名词	1193	3.438	n,名词	6029	3.602
v,动词	3938	2.625	v,动词	1123	3.236	v,动词	5772	3.448
w,标点符号	**1727**	**1.151**	**w,标点符号**	**809**	**2.331**	**w,标点符号**	**4893**	**2.923**
ns,地名	1520	1.013	ns,地名	312	0.899	d,副词	1593	0.952
vn,名动词	912	0.608	d,副词	249	0.718	**m,数词**	**1574**	**0.940**
m,数词	**733**	**0.489**	j,简称	185	0.533	**r,代词**	**1408**	**0.841**
d,副词	627	0.418	**m,数词**	**174**	**0.501**	u,助词	1161	0.694
a,形容词	472	0.315	nr,人名	135	0.389	q,量词	1066	0.637
q,量词	439	0.293	p,介词	132	0.380	nr,人名	964	0.576
p,介词	401	0.267	a,形容词	126	0.363	a,形容词	916	0.547
j,简称	373	0.249	u,助词	122	0.352	ns,地名	904	0.540
t,时间词	371	0.247	q,量词	122	0.352	p,介词	672	0.401
u,助词	269	0.179	**r,代词**	**117**	**0.337**	t,时间词	558	0.333
r,代词	**192**	**0.128**	vn,名动词	85	0.245	f,方位词	339	0.203
nr,人名	191	0.127	t,时间词	65	0.187	vn,名动词	323	0.193

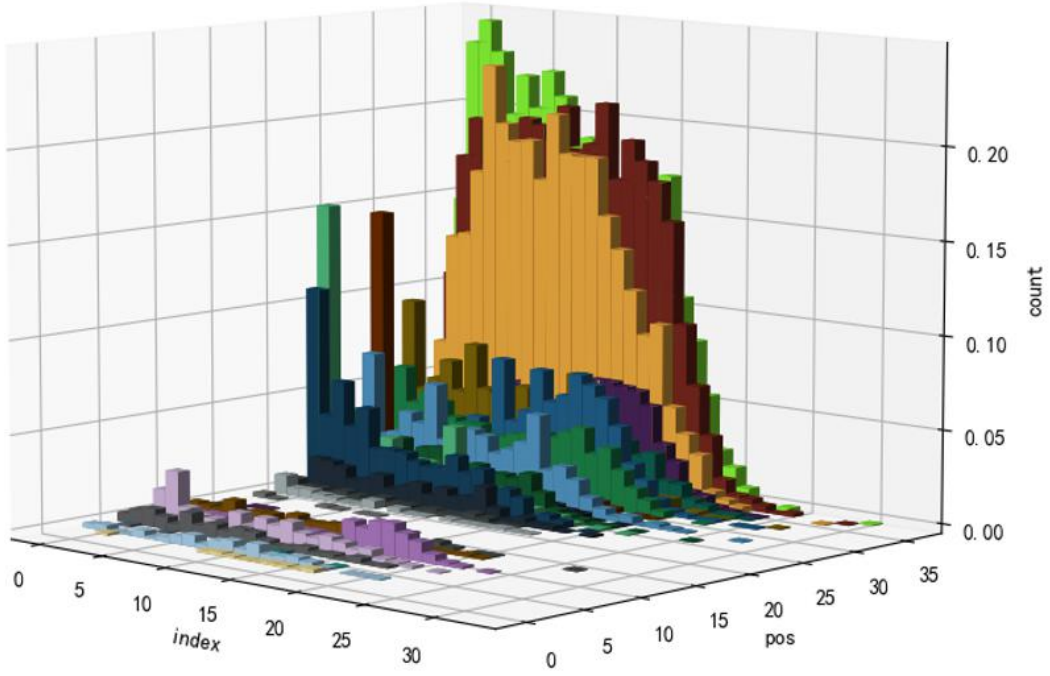
基于中文分词对比分析网络新闻标题

新闻标题结构对比分析

央视网



头条新闻



结论

偏严肃传统媒体（如央视网）

1. 更注重点明新闻内容
2. 核心词汇更多在标题前部出现

新型网络媒体（如头条新闻）

1. 拟标题不够简练，结构无定式
2. 大量使用标点符号，文风不严肃
3. 大量使用代词，指代可能不明

谢谢各位专家
敬请批评指正

Thanks for your listening

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY