北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

中文分词和词性标注

李佐瓯 章芷蕙 蒲沅东 李翰东

2022.3.11

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

1

中文分词

演讲：李佐瓯

中文分词：

中文的词通常由几个连续的字组成，词与词之间一般没有空格。我们把连续的字序列，按照一定的规范，重新组合成语义独立词序列的过程称之为分词。

例：

输入：亲 请问有什么可以帮您的吗?
输出：亲 / 请问 / 有 / 什么 / 可以 / 帮 / 您 / 的 / 吗 / ?

输入：共同创造美好的新世纪——二〇〇一年新年贺词
输出：共同 / 创造 / 美好 / 的 / 新 / 世纪 / —— / 二〇〇一年 / 新年 / 贺词

输入：因果关系不能假设但又必须假设
输出：因果 / 关系 / 不能 / 假设 / 但 / 又 / 必须 / 假设

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

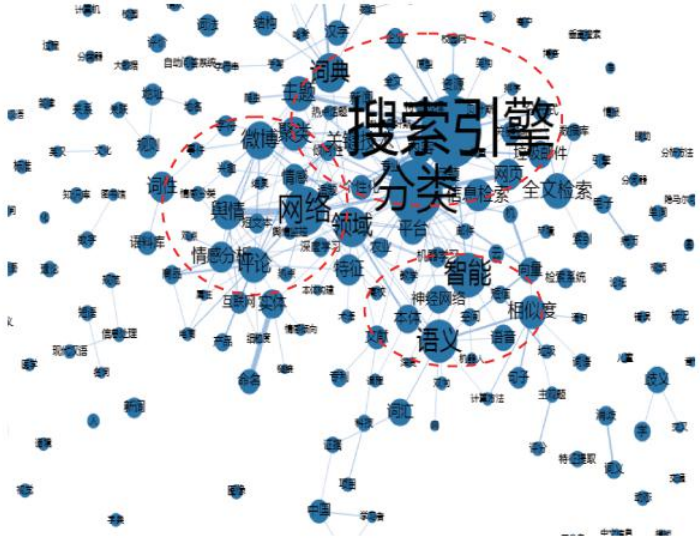
研究目的

- 中文分词是NLP中的基础工作
- 分词任务是很多NLP任务的前置

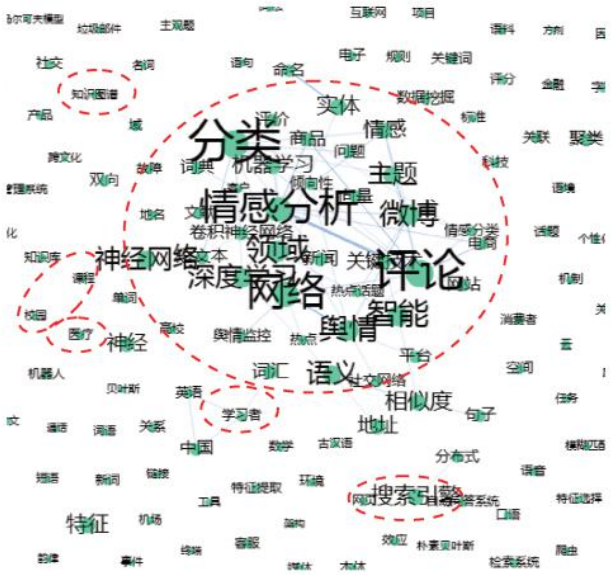
  词法分析、句法分析、语义分析、篇章分析等

- 分词任务有很多应用领域和场景

  搜索引擎、信息检索、舆情和情感分析、智能客服等



(a) 1984年–2019年



(b) 2017年–2019年

主要问题

➢分词标准不清晰，不统一，没有可计算的定义

中文词汇的开放性、动态性较高。词的标准因人而异，分词的标准也会因此不能统一。

SIGHAN-2005 数据集中就有约3%的切分不一致现象。

➢存在切分歧义

1. 交叉歧义(交集型切分歧义, OAS)， 覆盖歧义(多义组合型切分歧义, CAS)

   按时下的进展：”按时” / ”时下”　　有才能：”才” + ”能” / ”才能”

2. 真歧义，伪歧义

   签名球拍卖完了：签名球/拍卖/完/了　签名/球拍/卖完/了

➢未登录词(OOV)识别

新的通用词、专有名词、术语等 -> NER

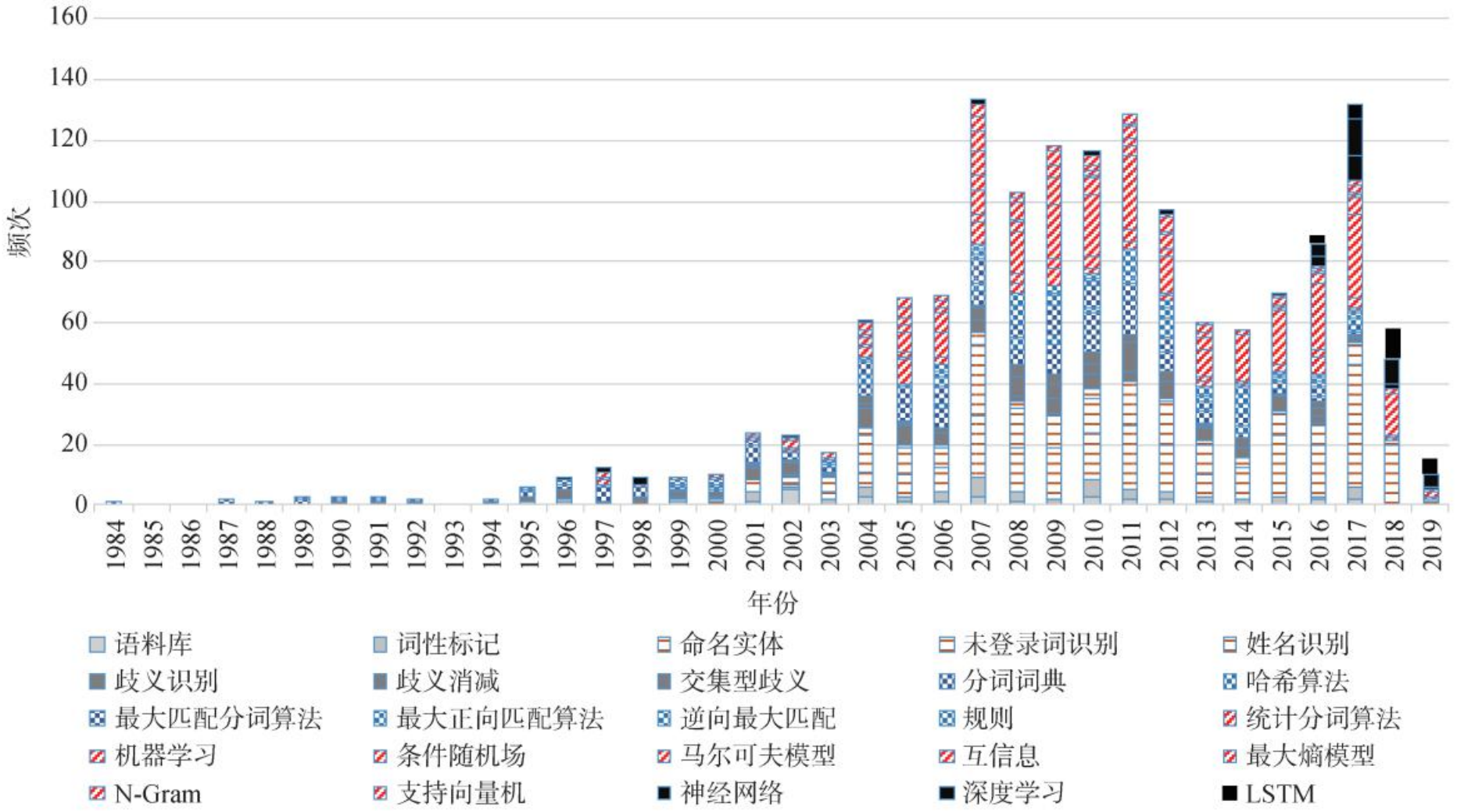


图 5 “中文分词文献”部分关键词分布(篇)

Fig.5 The Distribution of Key Words in Chinese Word Segmentation Literature

发展历程

机械分词

↓

机器学习

↓

深度学习

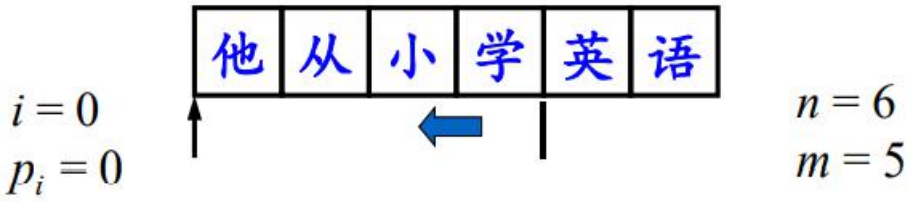
机械分词

➢基于词典的机械分词

正向最大匹配(Forward Maximum Matching, FMM)、逆向最大匹配(Reverse / Backward MM, FMM/BMM)
双向最大匹配、最短路径法、全切分法

- 输入：他从小学英语
- 词典：他，从小，从，小学，小，英语

(假设词典中最长词的长度为5)



FMM results： 他 从小 学 英语
BMM results： 他 从 小学 英语

例：正、逆向最大匹配

启发式规则：

1.如果正反向分词结果词数不同，则取分词数量较少的那个
2.如果分词结果词数相同
  a. 分词结果相同，就说明没有歧义，可返回任意一个
  b. 分词结果不同，返回其中单字较少的那个

例：双向最大匹配的一种启发式规则

机械分词

➢优点

1. 分词速度快

2. 汉语文本中90.0%左右的句子，FMM和BMM的切分完全重合且正确

➢存在的问题

1. 存在较多切分歧义问题

2. 分词准确率依赖于词典的好坏

3. 未登录词较多时效果较差

基于机器学习的分词

➢N-gram

假设第N个词的出现只与前N-1个词相关

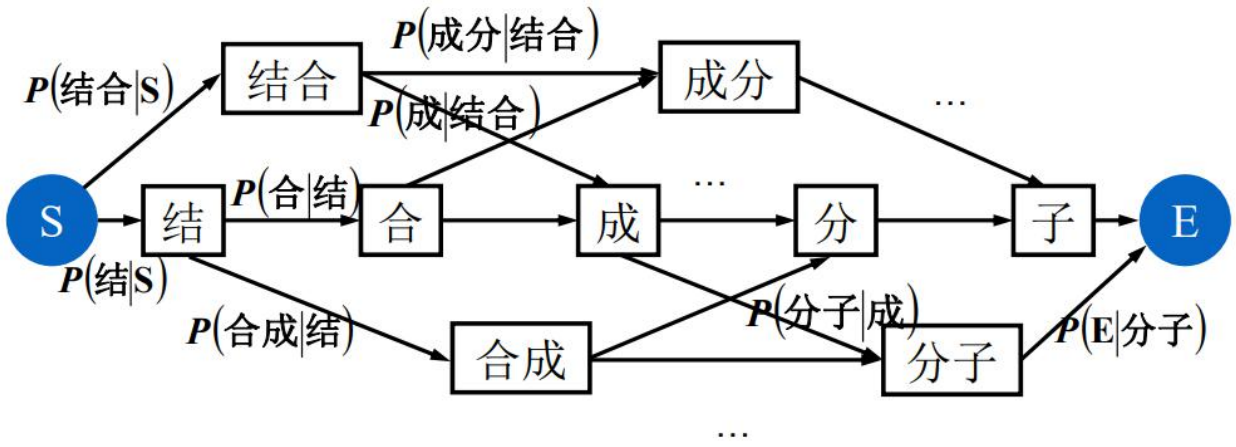
将文本里面的内容按照字节进行大小为N的滑动窗口操作

利用贝叶斯定理可推出

$$P(w_i|w_{i-n-1},\cdots,w_{i-1}) = \frac{C(w_{i-n-1},\cdots,w_i)}{C(w_{i-n-1},\cdots,w_{i-1})}$$

可以转化为在分词图上的最优路径搜索

无法识别未登录词



$$P(\text{结合 成 分子}) = P(\text{结合}|S)P(\text{成}|\text{结合})P(\text{分子}|\text{成})P(E|\text{分子})$$

例：2-gram

基于机器学习的分词

➢基于字标注　　使用2位或4位标记法 + 序列标注模型

迈/B向/I 充/B满/I 希/B望/I 的/B 新/B 世/B纪/I

迈/B向/E 充/B满/E 希/B望/E 的/S 新/S 世/B纪/E

隐马尔可夫模型(HMM)、最大熵马尔可夫模型(MEMM)、条件随机场(CRF)

条件随机场的使用最为广泛

1. 使用全局归一化避免偏移
2. 可以定义数量更多的特征函数
3. 可以使用任意的权重

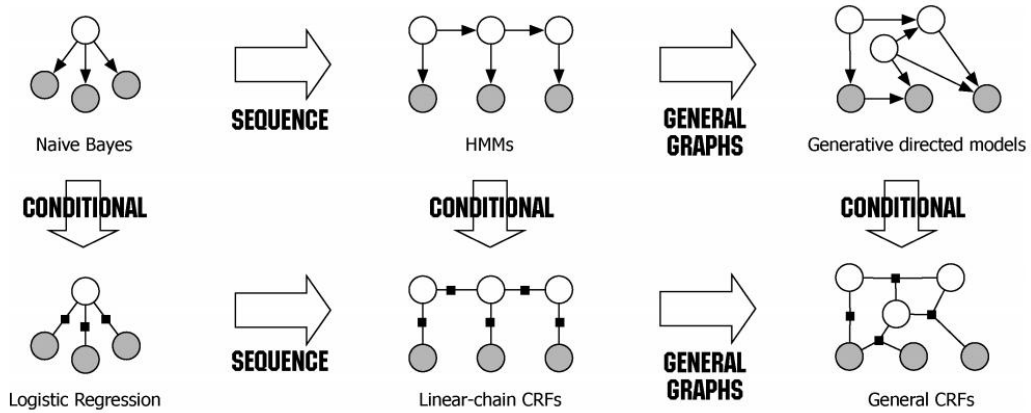


Fig. 2.4 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.

基于深度学习的分词

相较于机器学习方法

1. 无需进行人工的特征选择
2. 保留句子的长距离信息
3. 可以通过预训练等方法加入深层语言知识（句法、语用等）

通常的处理流程

1）字符嵌入层（unigram、bigram等）
2）Encoder用于提取上下文特征（CNN、Bi-LSTM、Transformer、BERT等）
3）Decoder用于进行标签推理（CRF、MLP(softmax)等）

数据集

State-of-the-art Chinese Word Segmentation with Bi-LSTMs

输入字和bigram，只使用：

1.简单的Bi-LSTM

2.预训练embedding： https://github.com/wlin12/wang2vec

3.在LSTM中添加Dropout

4.使用momentum-based averaged SGD（Weiss et al.2015）

和网格搜索进行调参就得到了很好的效果

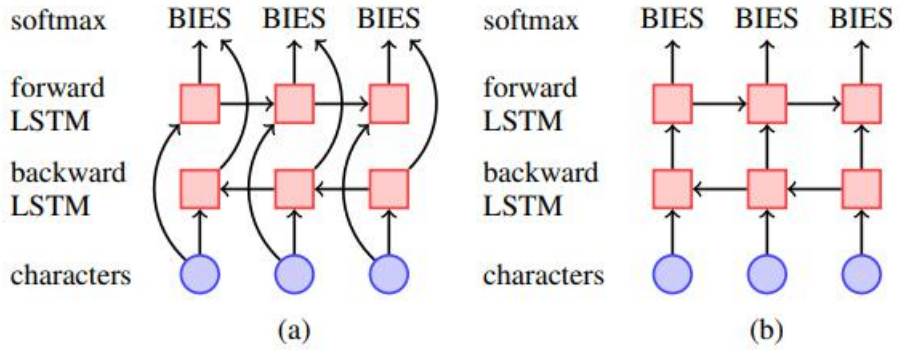


Figure 1: Bi-LSTM models: (a) non-stacking, (b) stacking. Blue circles are input (char and char bigram) embeddings. Red squares are LSTM cells. BIES is a 4-way softmax.

错误分析：

2/3的错误来自于未登录词，预训练embedding在遇到未登录词时可以提高10%的召回率
1/3的错误来自于数据集的标注错误，标注的不一致性会影响模型在该数据集上的表现

State-of-the-art Chinese Word Segmentation with Bi-LSTMs (Ma et al., EMNLP 2018)

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Glyce: Glyph-vectors for Chinese Character Representations

1.创新地将字形信息融入到中文NLP中

2.针对于12x12大小的字形图片，设计了用于汉字图像处理的Tianzige-CNN架构

3.把汉字图像分类加入多任务训练中，减少过拟合现象

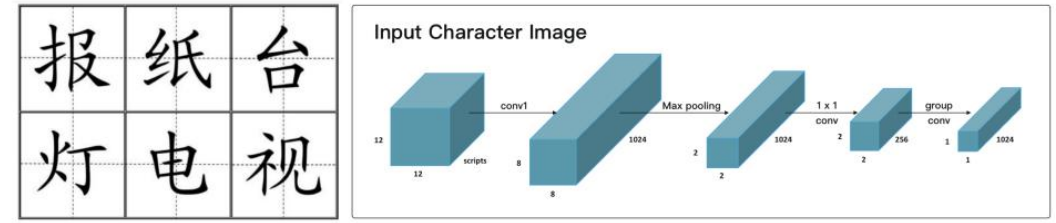


Figure 1: Illustration of the Tianzege-CNN used in Glyce.

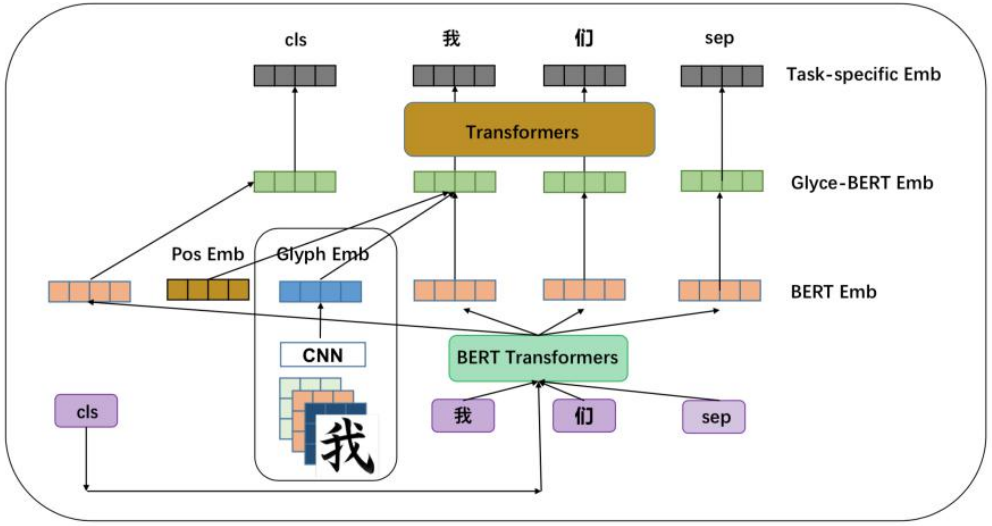


Figure 2: Combing glyph information with BERT.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, Jiwei Li: "Glyce: Glyph-vectors for Chinese Character Representations", 2019; arXiv:1901.10125

Towards Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning

从分词标准的不固定出发，试图学习所有分词标准的共性以及其 underlying knowledge

1. 使用BERT进行上下文特征提取
2. 使用量化、剪枝、编译器优化等技巧提高计算效率
3. 在BERT后增加一个域映射层，其中的私有层用来提取标准特定的特征，公有层用来提取公共特征
4. 对于每个标准，将私有层的隐藏状态和公共层隐藏状态拼接，输入到CRF中，计算得分函数。

从多种分词标准中学习可以有效缓解歧义带来的分词错误

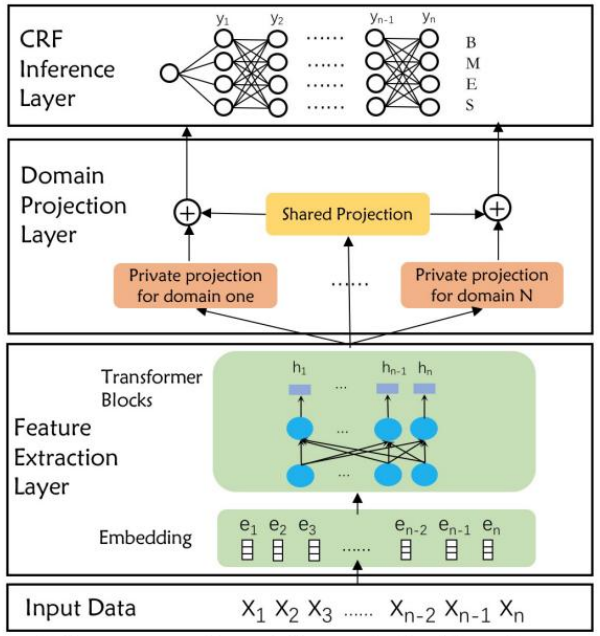


Figure 1: The architecture of the proposed model, stacked with a feature extraction layer, a domain projection layer and a CRF tag inference layer.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, Wei Chu: "Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder

共享来自多个分词标准中的知识，可以应对多标准中文分词任务

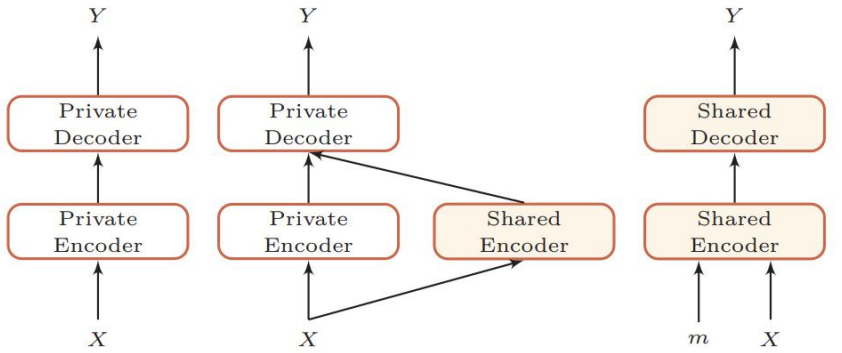
在模型中：

1. 共享编码器用于抽取对分词标准敏感的语境特征（criteria-aware contextual features）
2. 共享解码器则用于预测针对标准而不同的标签（criteria-specific labels）。

在MSRA, AS, PKU, CTB等多个数据集中，
P, R, F, OOV指标都达到SOTA

Table 1: Illustration of the different segmentation criteria.

Corpora	Lin	Dan	won	the championship	
CTB	林丹		赢得	总冠军	
PKU	林	丹	赢得	总	冠军
MSRA	林丹		赢得	总	冠军



(a) Single criterion CWS (b) MTL-based Multi-criteria CWS (c) Transformer-based Multi-criteria CWS

Figure 1: Architectures of single-criterion and multi-criteria Chinese word segmentation. The red components are shared.

Xipeng Qiu, Hengzhi Pei, Hang Yan, Xuanjing Huang: "A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder", 2019;

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Improving Chinese Word Segmentation with Wordhood Memory Networks

通过增加 Memory Networks 缓解OOV问题

1. Encoder 使用 BERT/ZEN
2. Decoder 使用 MLP 或者 CRF
3. 通过当前词的表示 + N-gram和Label Value 的 embedding 做Softmax 得到相关性

Memory Networks 可以对N-gram进行语义信息匹配求最大，从而学习哪些N-gram的形成对于最后完整表达句意的帮助更大来提升效果。

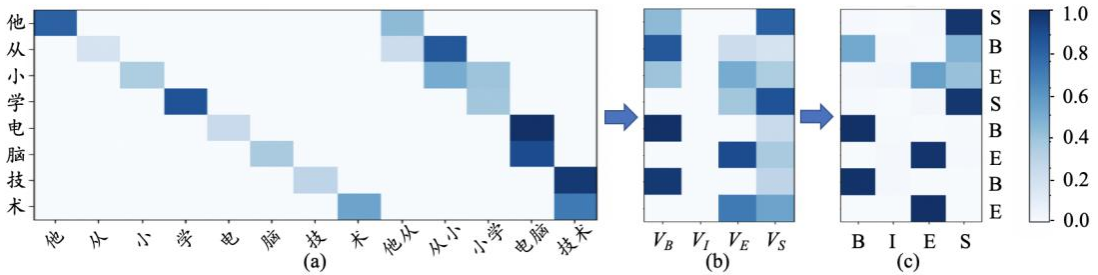


Figure 3: Heatmaps of weights learned for (a) keys and (b) values in the memory, and (c) the tags from the decoder, with respect to each character in an input sentence. Higher weights are visualized with darker colors.

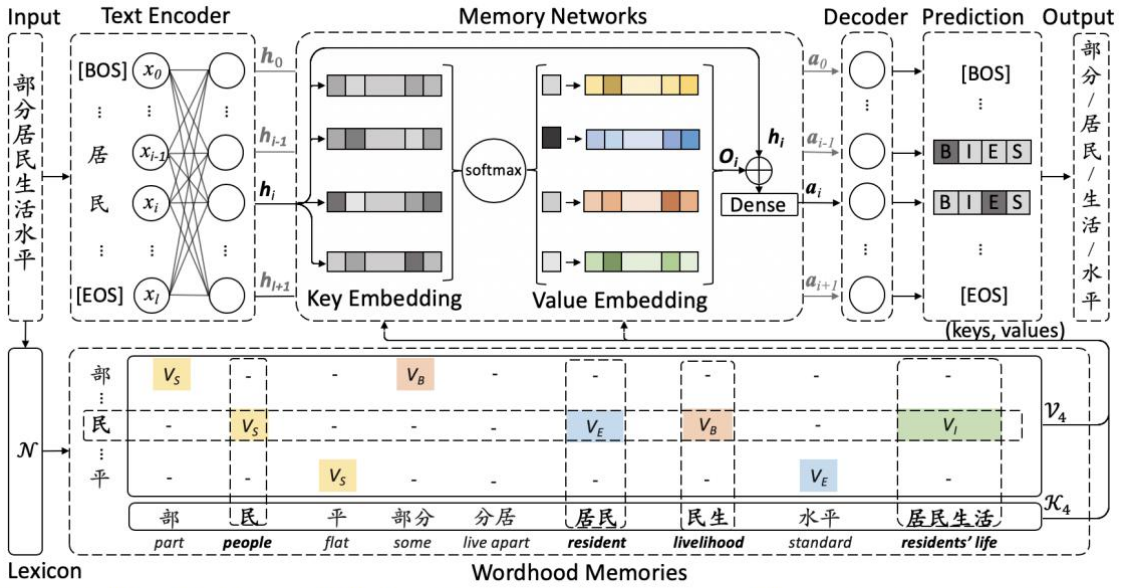


Figure 1: The architecture of WMSEG. "$\mathcal{N}$" denotes a lexicon constructed by wordhood measures. N-grams (keys) appearing in the input sentence "部分居民生活水平" (*some residents' living standard*) and the wordhood information (values) of those n-grams are extracted from the lexicon. Then, together with the output from the text encoder, n-grams (keys) and their wordhood information (values) are fed into the memory module, whose output passes through a decoder to get final predictions of segmentation labels for every character in the input sentence.

Improving Chinese Word Segmentation with Wordhood Memory Networks (Tian et al., ACL

改进方向

集成算法

深度学习方法 Encoder + 统计学习方法 Decoder + 规则方法(后处理)

联合模型

分词+词性标注、分词+词性标注+句法分析、分词+句法分析

蒸馏

分词任务通常需要较高的性能，如何把 大BERT 变成性能相近的 小CNN/LSTM/BERT ？

多粒度、标准分词

如何通过不同粒度标准的分词语料联合预训练，让分词器通过简单的控制
适应不同的分词场景?

Sproat R, Shih C. A Statistical Method for Finding Word Boundaries in Chinese Text[J]. Computer Processing of Chinese and Oriental Languages, 1990, 4(4): 336-351.

Huang C N, Zhao H. Which is Essential for Chinese Word Segmentation: Character Versus Word[C]// Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, Wuhan, China. Beijing, China: Tsinghua University Press, 2006: 1-12.

Xue N. Chinese Word Segmentation as Character Tagging[J]. Computational Linguistics & Chinese Language Processing, 2003, 8(1): 29-47.

Berger A L, Pietra V J D, Pietra S A D. A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistics, 1996, 22(1): 39-71.

Rabiner L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

McCallum A, Freitag D, Pereira F C N. Maximum Entropy Markov Models for Information Extraction and Segmentation [C]// Proceedings of the 17th International Conference on Machine Learning, CA, USA. CA, USA: ICMS, 2000.

Peng F, Feng F, McCallum A. Chinese Segmentation and New Word Detection Using Conditional Random Fields[C]// Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland. New York, USA: ACL, 2004.

Tseng H, Chang P, Andrew G, et al. A Conditional Random Field Word Segmenter for SIGHAN Bakeoff 2005[C]// Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea. New York, USA: ACL, 2005

中文分词

Pre-training with Meta Learning for Chinese Word Segmentation (Ke et al., NAACL 2021)

A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder (Qiu et al., Findings 2020)

Improving Chinese Word Segmentation with Wordhood Memory Networks (Tian et al., ACL 2020)

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, Jiwei Li: "Glyce: Glyph-vectors for Chinese Character Representations", 2019; arXiv:1901.10125.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, Wei Chu: "Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning", 2019; arXiv:1903.04190.

State-of-the-art Chinese Word Segmentation with Bi-LSTMs (Ma et al., EMNLP 2018)

Neural Word Segmentation with Rich Pretraining (Yang et al., ACL 2017)

Word-Context Character Embeddings for Chinese Word Segmentation (Zhou et al., EMNLP 2017)

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（2/3）：1-17. (Tang Lin, Guo Chonghui, Chen Jingfeng. Review of Chinese Word Segmentation Studies[J]. Data Analysis and Knowledge Discovery, 2020, 4(2/3): 1-17.)

https://chinesenlp.xyz/#/docs/word_segmentation

https://www.zhihu.com/question/19578687

鉴萍，黄河燕．《自然语言处理》 2021-2022 (1)

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

2

中文词性标注

演讲：章芷蕙

中文词性标注是在语料库中根据词的定义和上下文，用适当的词类：动词、名词、形容词等标注词的过程。

中文词性标注序列，标签序列，词性序列示例

句子序列	北京	是	一	座	包容	的	城市
标签序列	n	v	m	q	a	u	n
词性序列	名词	动词	数词	量词	形容词	助词	名词

研究难点：

- 缺乏词形态变化
- 常用词兼类现象严重
- 研究者主观原因
- 训练语料的缺失

研究意义：

- 能够在很大程度上消除词义歧义
- 提高句子检索性能
- 提高区分信息新颖程度的能力

词性标注发展历程

规则方法 → 1971年，美国布朗大学的Greene和Rubin建立了TAGGIT系统

统计方法 → 1983年，G.Leech和R.Garside等人建立了CLAWS系统

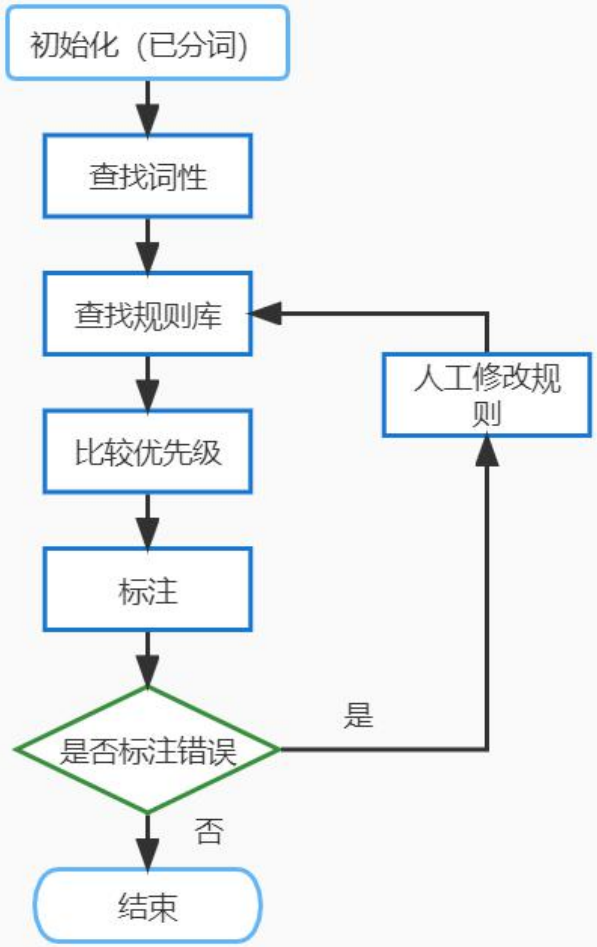
规则和统计融合方法 → 1996年，北京大学计算语言学研究所提出先规则、后统计的算法

深度学习方法 → 2011年，Collobert等人提出一种神经网络序列标注模型

基于规则的中文词性标注

《一种基于规则优先级的词性标注方法》

王广正等人[1]通过构建规则库，并根据中文语料库中各规则出现的频度调整每条规则的优先级



基于统计的中文词性标注

主要统计词性标注模型

- ME
- HMM
- CRF

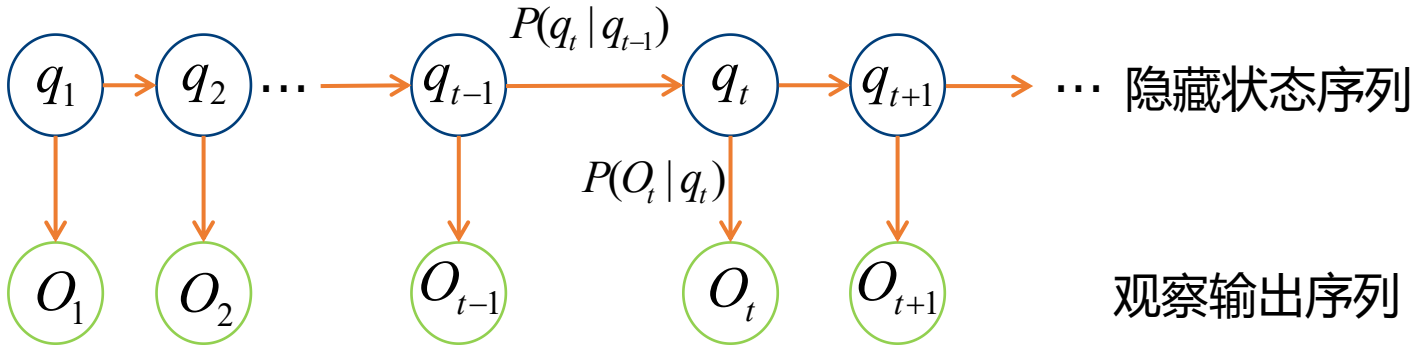
《Chinese POS tagging based on maximum entropy model》

[2]提出一种基于最大熵模型的中文词性标注方法，该方法挖掘并结合了所有有助于预测词性标注的特征。

基于统计的中文词性标注

《Hidden Markov Model: Unsupervised POS Tagging with Context Local Tag Generation》

[3]使用HMM模型将状态的子集分配给以局部域为条件的分布



HMM图解
隐状态是词性，显状态是单词

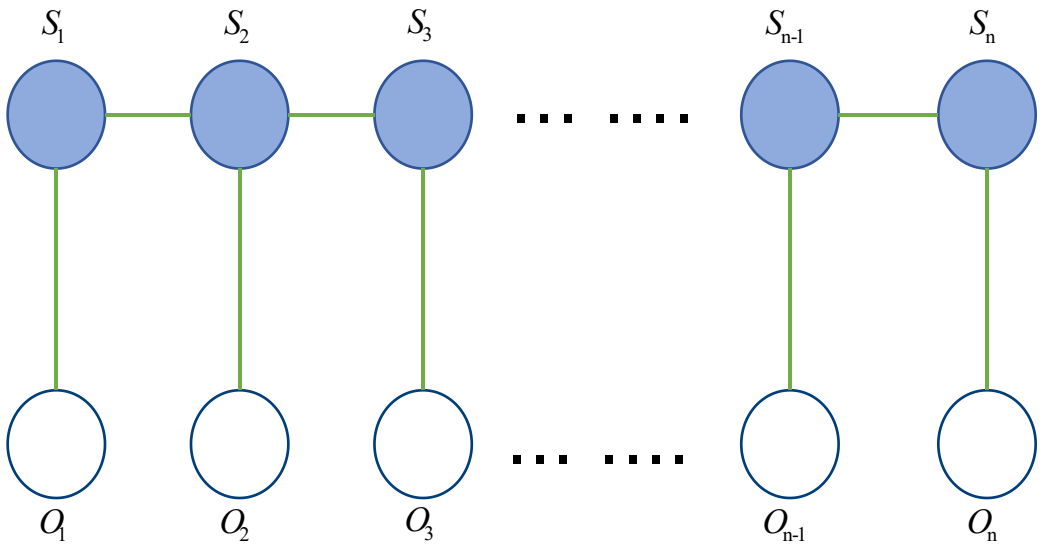
《基于半监督隐马尔科夫模型的汉语词性标注研究》

[4]使用一个小规模的训练语料库进行半监督隐马尔可夫学习

基于统计的中文词性标注

《基于条件随机场的汉语词性标注》

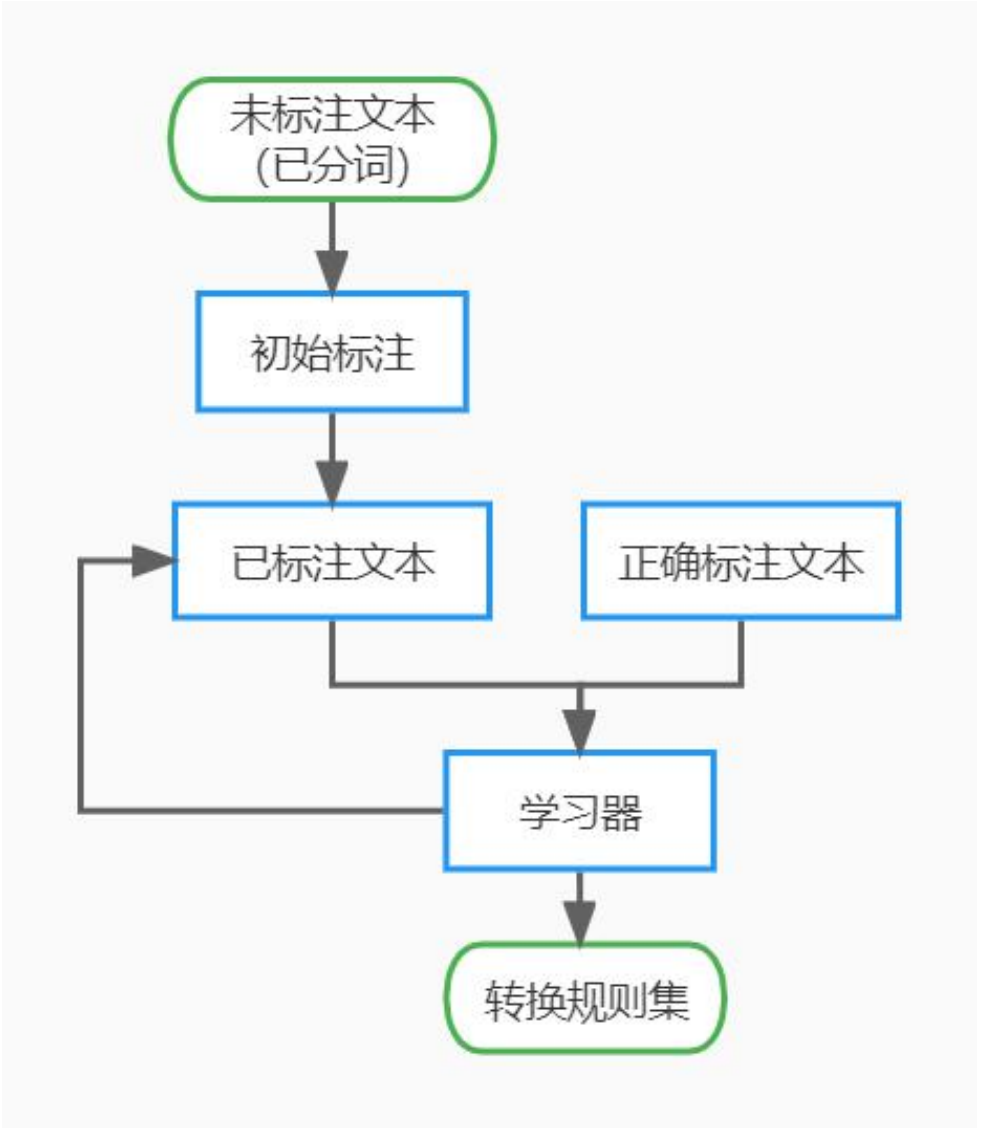
[5]利用CRF++工具包所采用的特征模板对CRF词性标注模型进行优化



CRF简单链图或线图

基于规则和统计的中文词性标注

将规则方法和统计方法相结合能够互相弥补不足之处，所以出现第三种方法:统计和规则相结合的词性标注方法。



基于规则和统计的中文词性标注

《A Method Integrating Rule and HMM for Chinese Part-of- speech Tagging》

[6]利用规则和HMM联合消除文本中的汉语词性歧义

《Part-of-speech Tagging Based on Dictionary and Statistical Machine Learning》

[7]分别基于规则和ME模型标注词语，并比较，若由规则方法获得的标签包含由ME方法获得的标签，则选择ME中的标注作为最终标注

基于深度学习的中文词性标注

严重依赖人工处理

规则、统计方法 → 最终得到的结果不稳定

深度学习方法

常用模型：RNN、LSTM、CNN和GRU等。

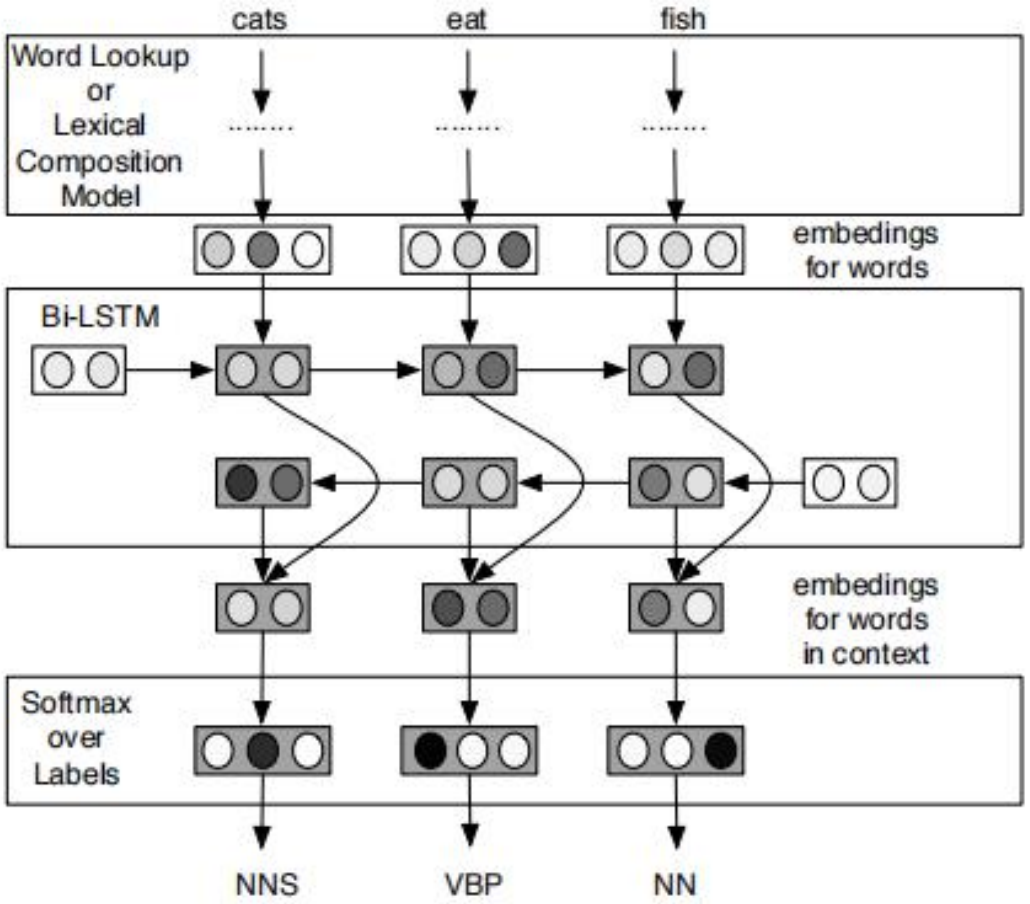
无法跟踪句子中词之间的长期关系

对数据多层建模
自动获取特征信息

基于深度学习的中文词性标注

《Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation》

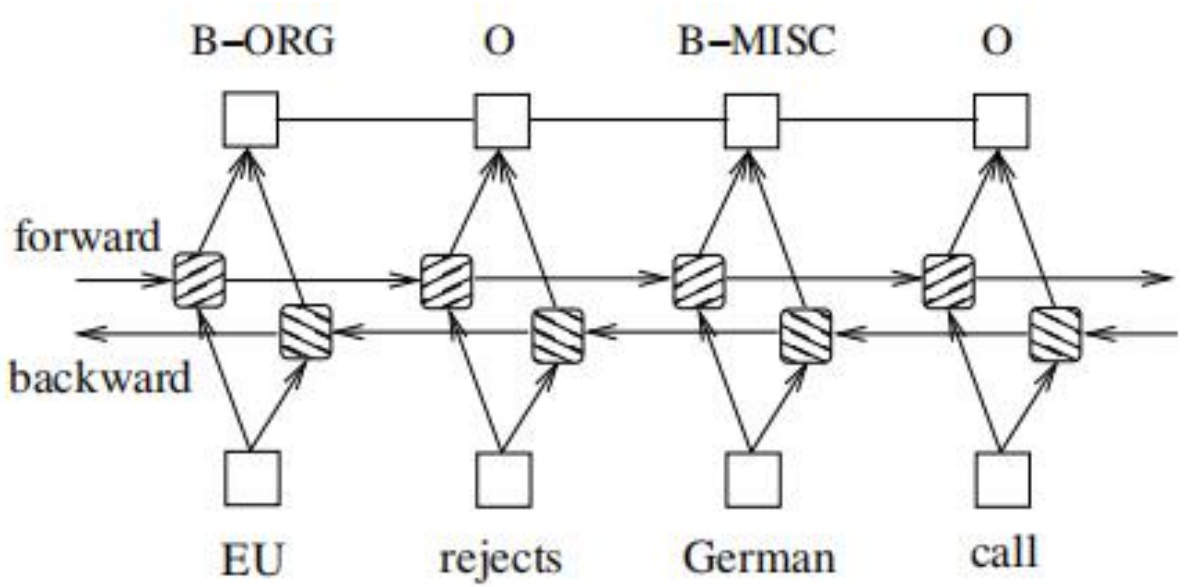
[8]2015年，提出C2W模型，通过双向LSTM组合字符来构建词向量



基于深度学习的中文词性标注

《Bidirectional LSTM-CRF Models for Sequence Tagging》

[9]将CRF与双向LSTM结合，LSTM层有效地使用过去的输入特征，CRF层使用句子级标记信息

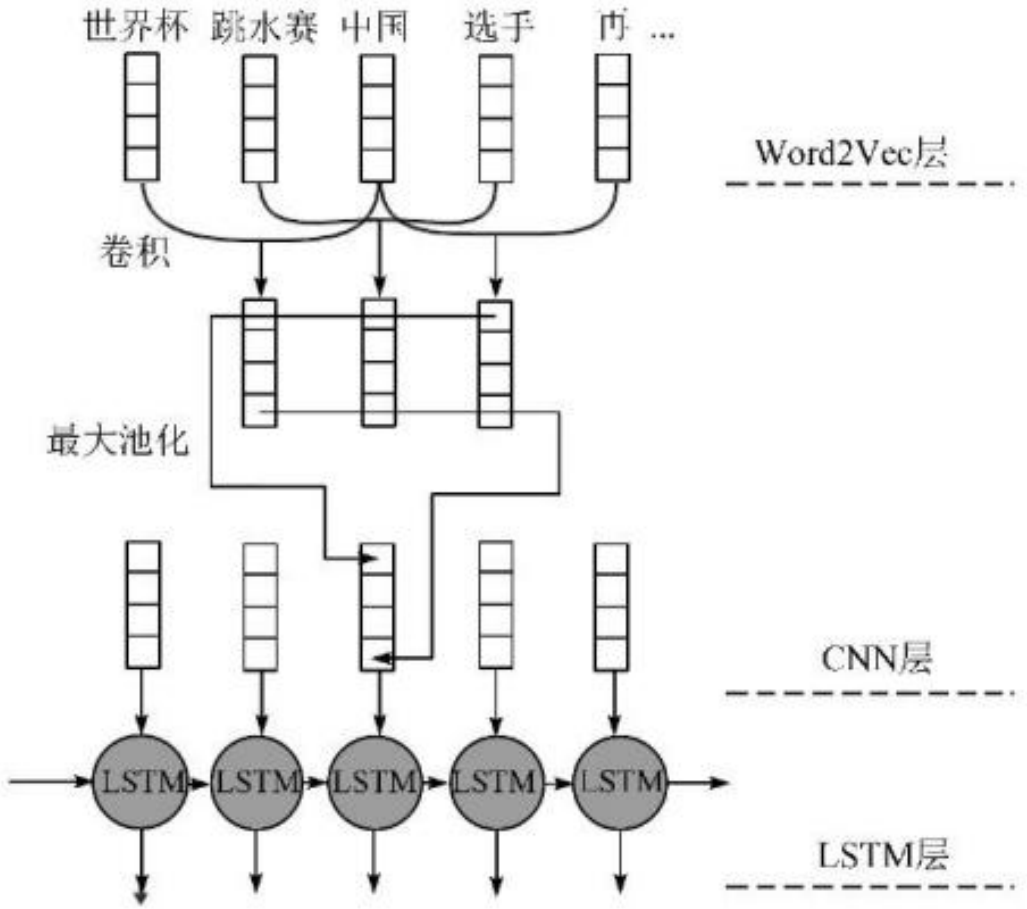


北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

基于深度学习的中文词性标注

《基于CNN和LSTM混合模型的中文词性标注简》

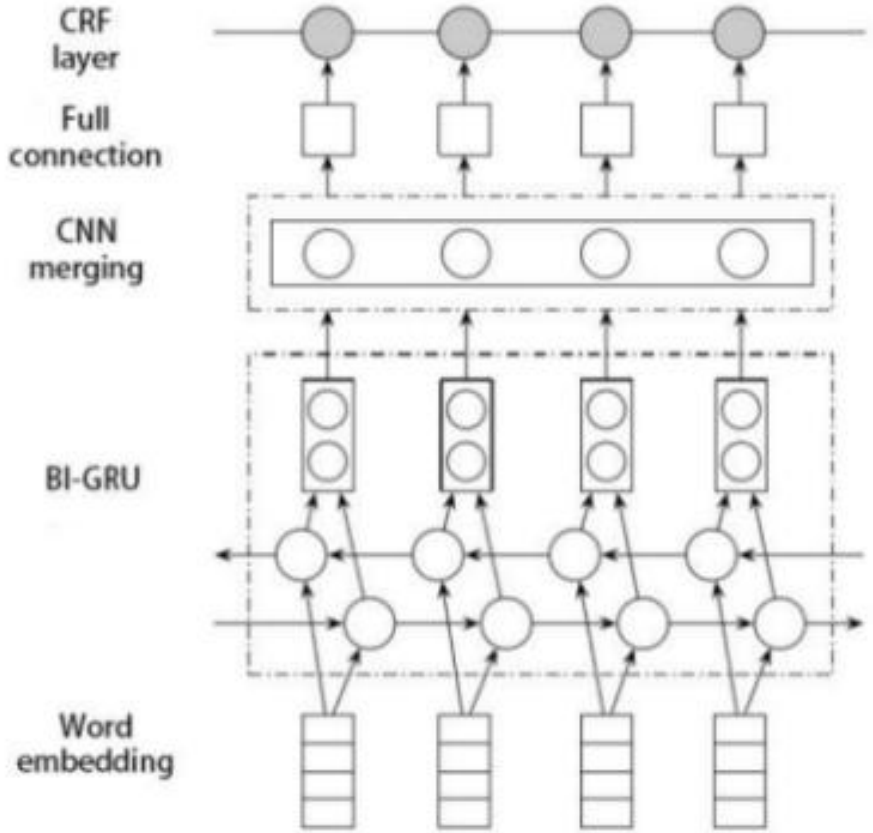
[10]2017年，利用CNN滑动窗口的特性获取词语表示特征，并通过LSTM来产生词性标注的序列标签



基于深度学习的中文词性标注

《Sequence Labeling of Chinese Text Based on Bidirectional Gru-Cnn-Crf Model》

[11]2018年，刘等人提出一种基于双向GRU-CNN-CRF的混合中文序列标注模型

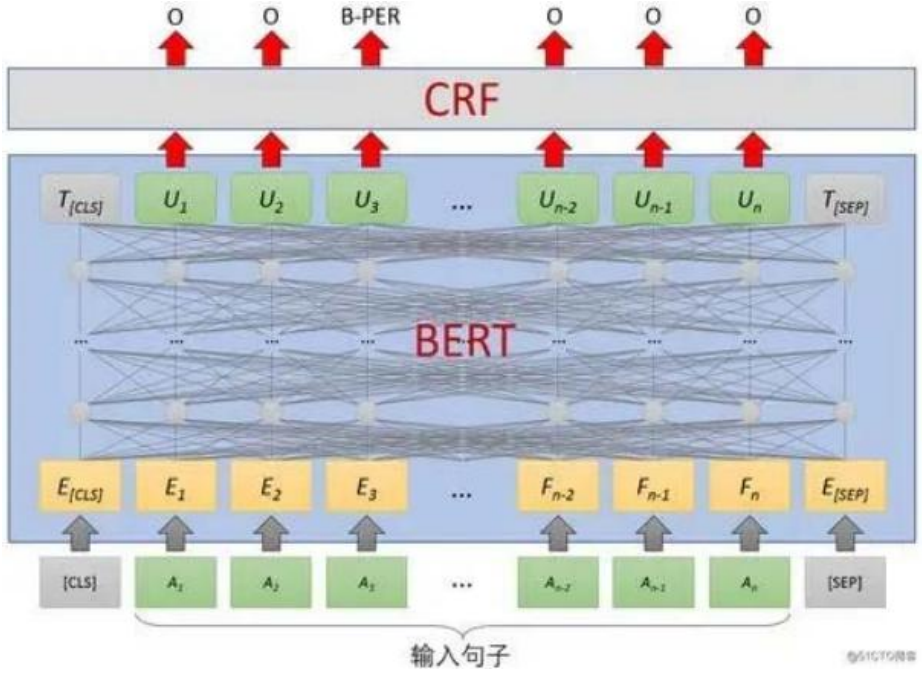


前沿进展及展望

当下热门技术如Attention、BERT等是目前研究的主流方向，适用于中文词性标注，相较于传统的神经网络有较大的优势。

提高正确率和消歧率

加快标注速度



参考文献

[1]王广正，王喜凤．一种基于规则优先级的词性标注方法[J]．安徽工业大学学报：自然科学版，2008，25(4):4
[2]Zhao J，Wang X L. Chinese POS tagging based on maximum entropy model[C]// International Conference on Machine Learning & Cybernetics. IEEE，2002.
[3]Moon T，Erk K，Baldridge J. Crouching Dirichlet，Hidden Markov Model: Unsupervised POS Tagging with Context Local Tag Generation[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing，EMNLP 2010，9-11 October 2010，MIT Stata Center，Massachusetts，USA，A meeting of SIGDAT，a Special Interest Group of the ACL. Association for Computational Linguistics，2010.
[4]韩霞，黄德根．基于半监督隐马尔科夫模型的汉语词性标注研究[J].小型微型计算机系统，2015,36(12):4.
[5]于江德，葛彦强，余正涛．基于条件随机场的汉语词性标注[C]// CNKI. CNKI，2011:63-66.
[6]Hui N，Hua Y，Li Z. A Method Integrating Rule and HMM for Chinese Part-of- speech Tagging[C]// 2007 2nd IEEE Conference on Industrial Electronics and Applications. IEEE，2007.
[7]YE，Zhonglin，JIA，et al. Part-of-speech Tagging Based on Dictionary and Statistical Machine Learning[C]// IEEE 2016 35th Chinese Control Conference (CCC) - Chengdu，China (2016.7.27-2016.7.29)
[8]Ling W，T Luís，Marujo L，et al. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation[J]. Computer Science，2015:1899-1907.
[9]Huang Z，Wei X，Kai Y. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science，2015.
[10]谢逸，饶文碧，段鹏飞,等．基于CNN和LSTM混合模型的中文词性标注简[J]．武汉大学学报：理学版，2017，63(3):5.
[11]Liu D，X Zou. Sequence Labeling of Chinese Text Based on Bidirectional Gru-Cnn-Crf Model[C]// 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE，2019.

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

3

中文分词和标注联合模型

演讲：蒲沅东

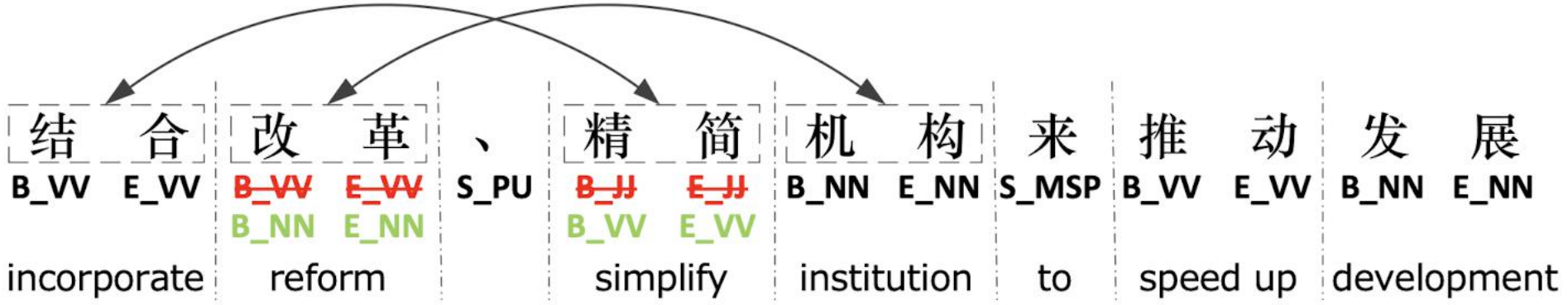
Why Joint Models?

Method	Word Seg F-measure (%)	POS Accuracy (%)	Total Testing Time
One-at-a-Time Word-Based	95.1	84.1	1 min 20 secs
One-at-a-Time Char-Based	95.1	91.7	1 min 50 secs
All-At-Once Char-Based	95.2	91.9	20 mins

[1]Chinese Part-of-Speech Tagging:
One-at-a-Time or All-at-Once?
Word-Based or Character-Based?

1.防止误差传播
2.利用标注信息来帮助分词

How Joint Models?



- *b*: the begin of the word
- *m*: the middle of the word
- *e*: the end of the word
- *s*: a single-character word

一种特殊的标注任务

问题：
1.搜索空间大
2.计算量大

What Joint Models?

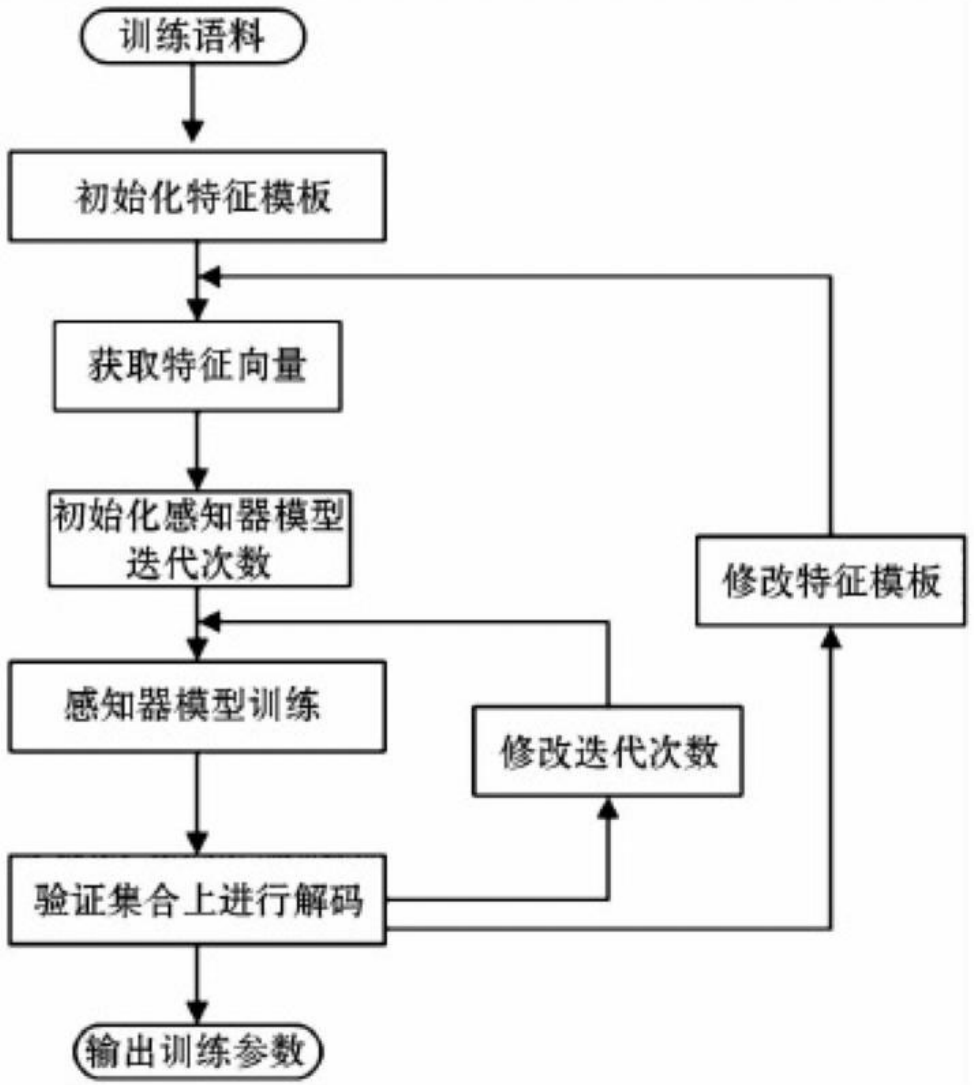
传统模型

深度学习模型

传统联合模型

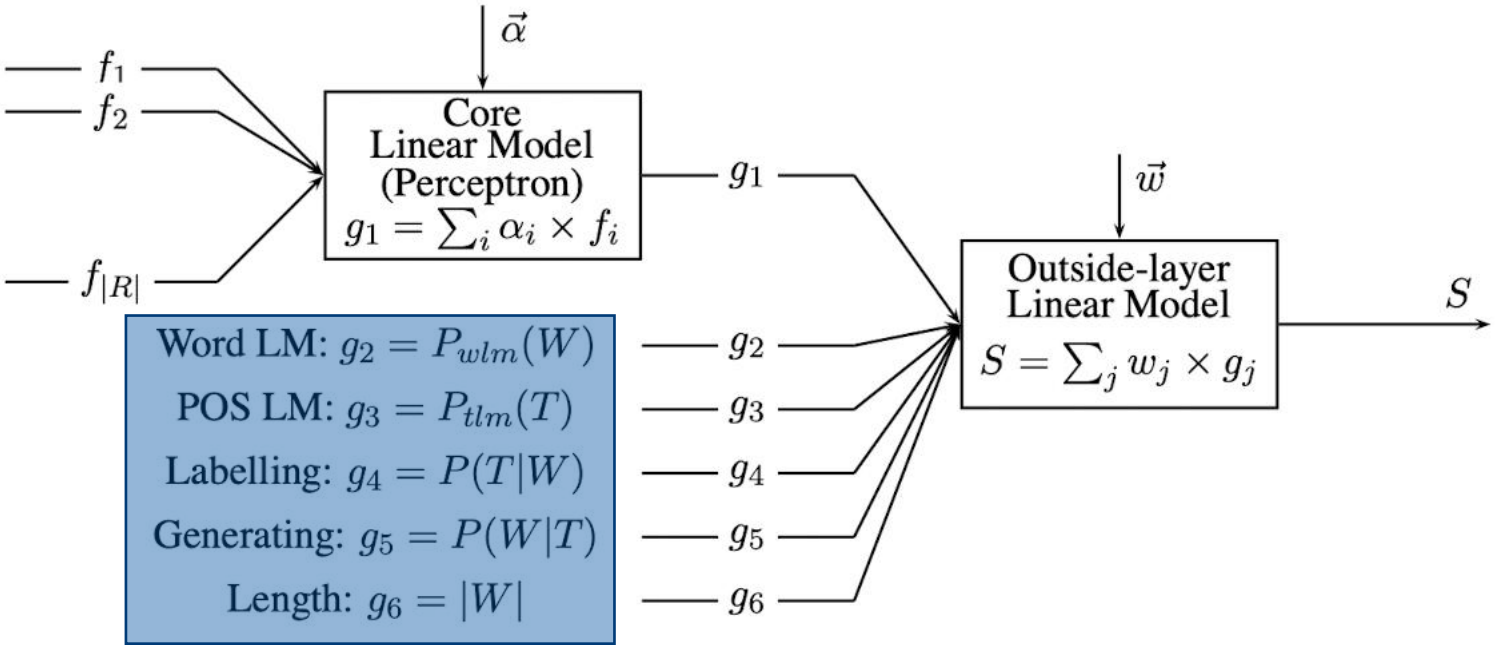
对句子的基本处理单位?

- 基于字的联合模型
- 基于词的联合模型
- 混合联合模型

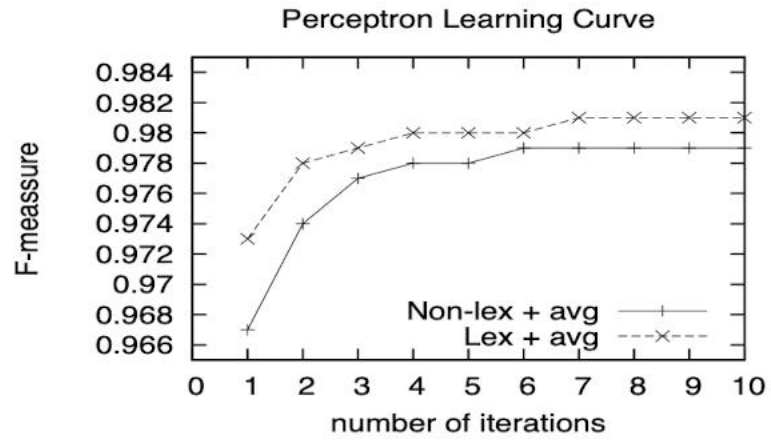


传统联合模型

·基于字的联合模型



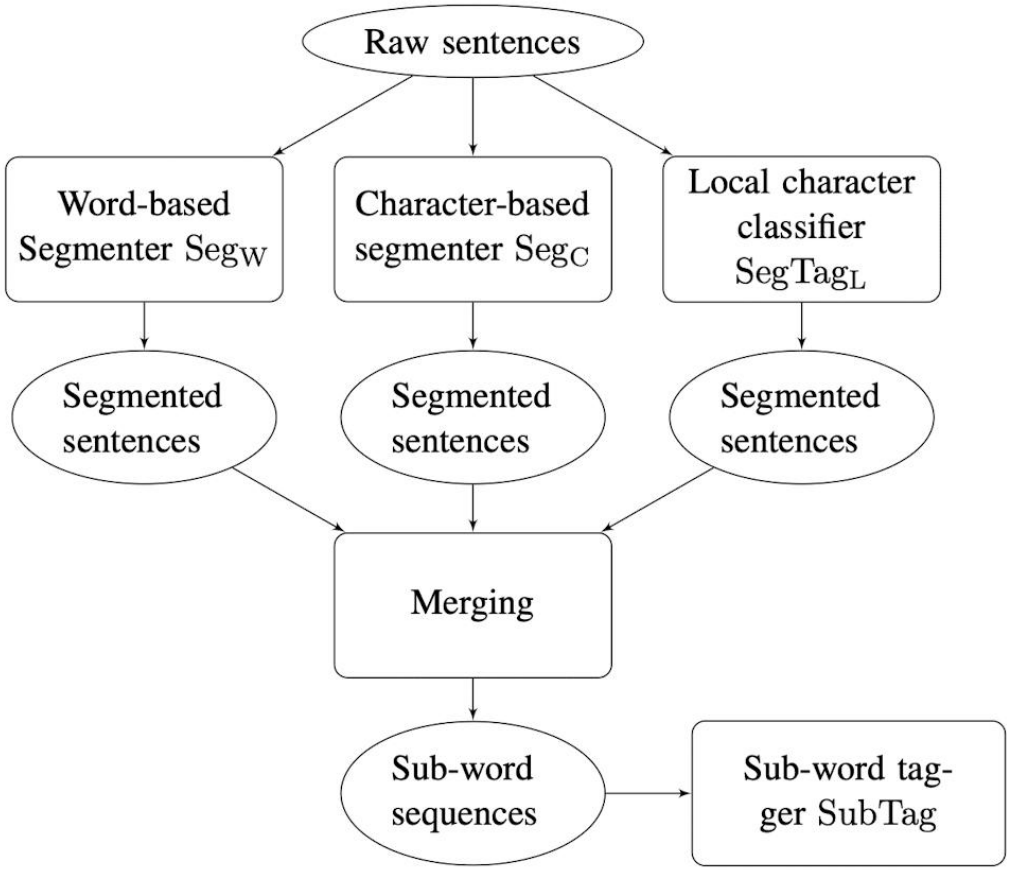
	AS	CityU	PKU	MSR
SIGHAN best	0.952	0.943	0.950	0.964
Zhang & Clark	0.946	0.951	0.945	0.972
our model	0.954	0.958	0.940	0.975



【2】A cascaded linear model for joint chinese word segmentation and part-of-speech tagging

传统联合模型

·混合联合模型



Test	Seg	Seg&Tag
(Jiang et al., 2008a)	97.85	93.41
(Jiang et al., 2008b)	97.74	93.37
(Kruengkrai et al., 2009)	97.87	93.67
(Zhang and Clark, 2010)	97.78	93.67
Our system	**98.17**	**94.02**

Table 7: F-score performance on the test data.

【3】A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging

深度学习模型

传统模型的缺点

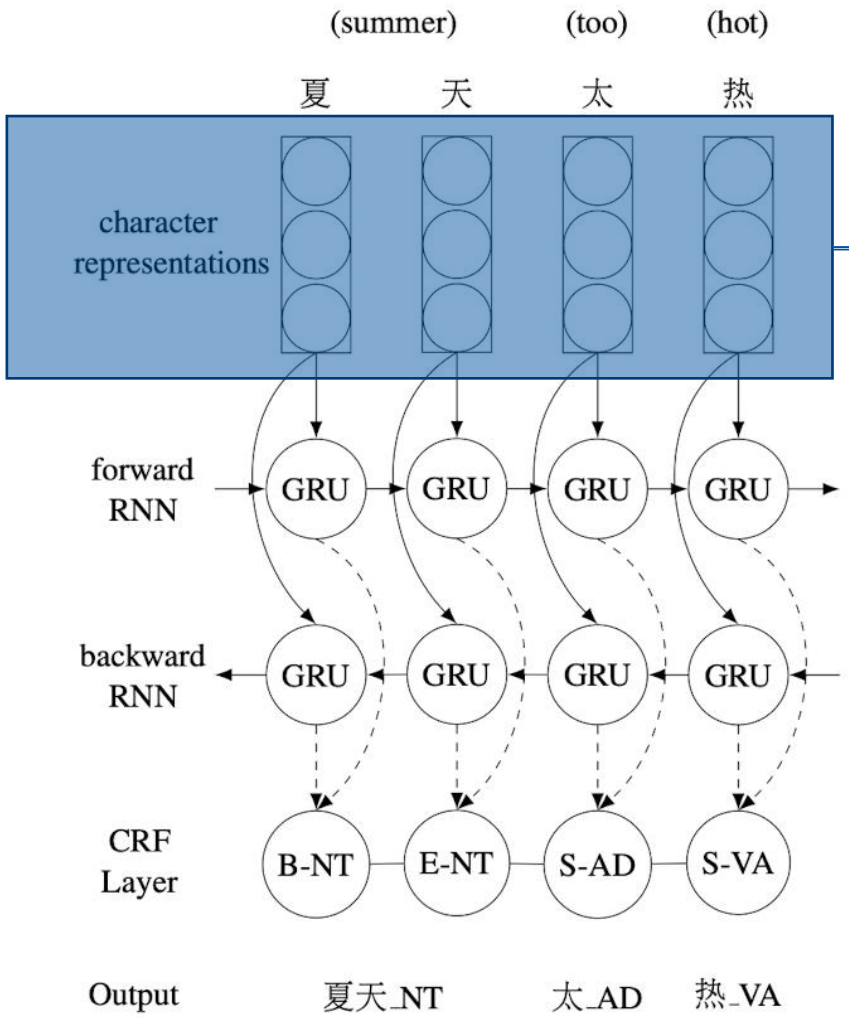
1.传统模型过大导致计算和存储困难
2.传统模型参数过多容易过拟合
3.搜索空间过大，难以解码

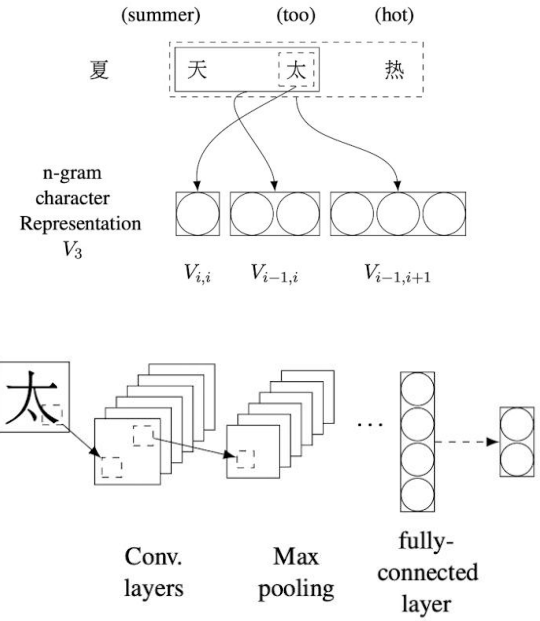
Deep Learning!

深度学习模型

	Sequence-labeling models	Transition-based models
特点	可以使用动态解码	能够更加灵活的进行特征工程

BiRNN(GRU)+CRF[4]

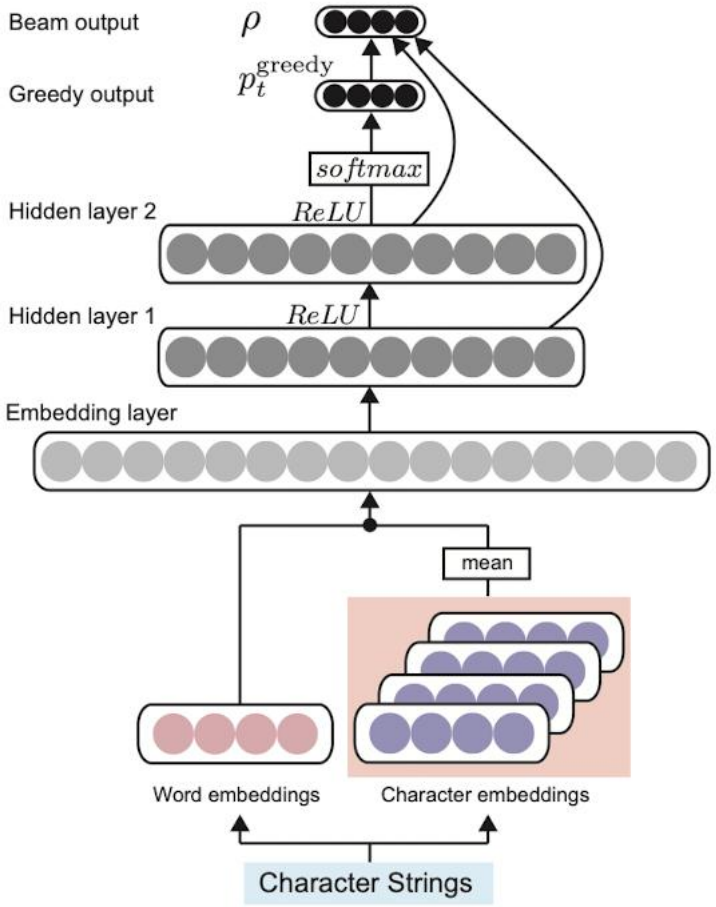


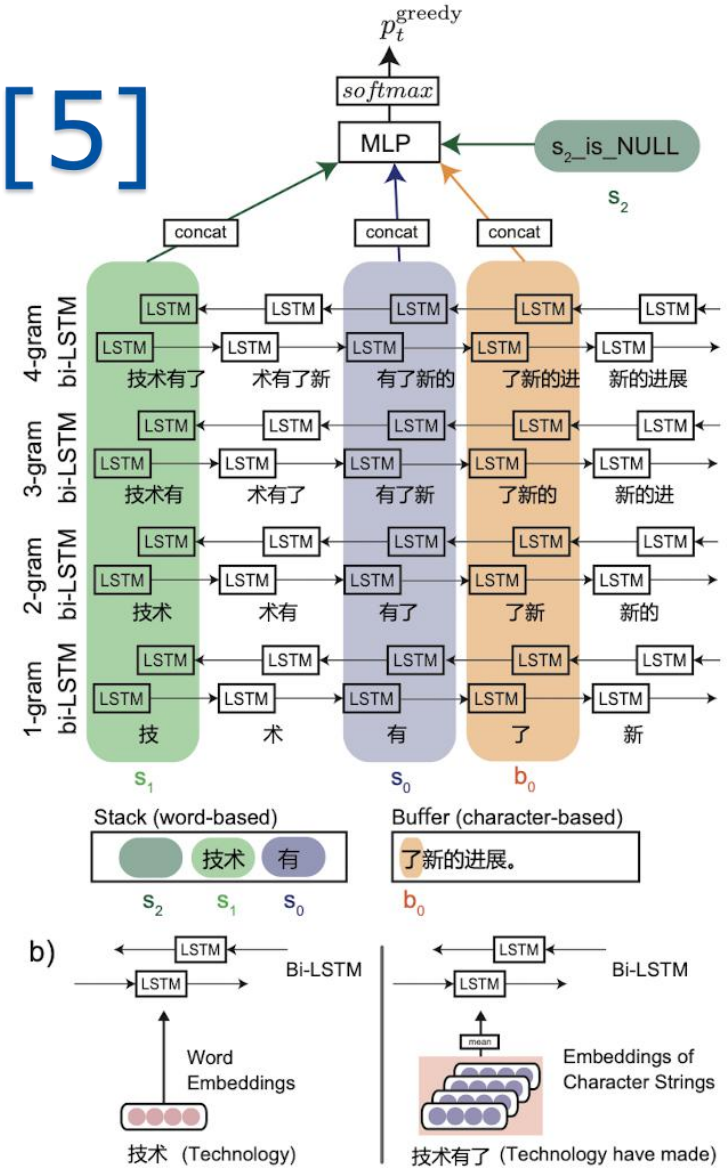


预训练字符 Embedding

偏旁部首 Embedding

Feed-forward&bi-LSTM[5]





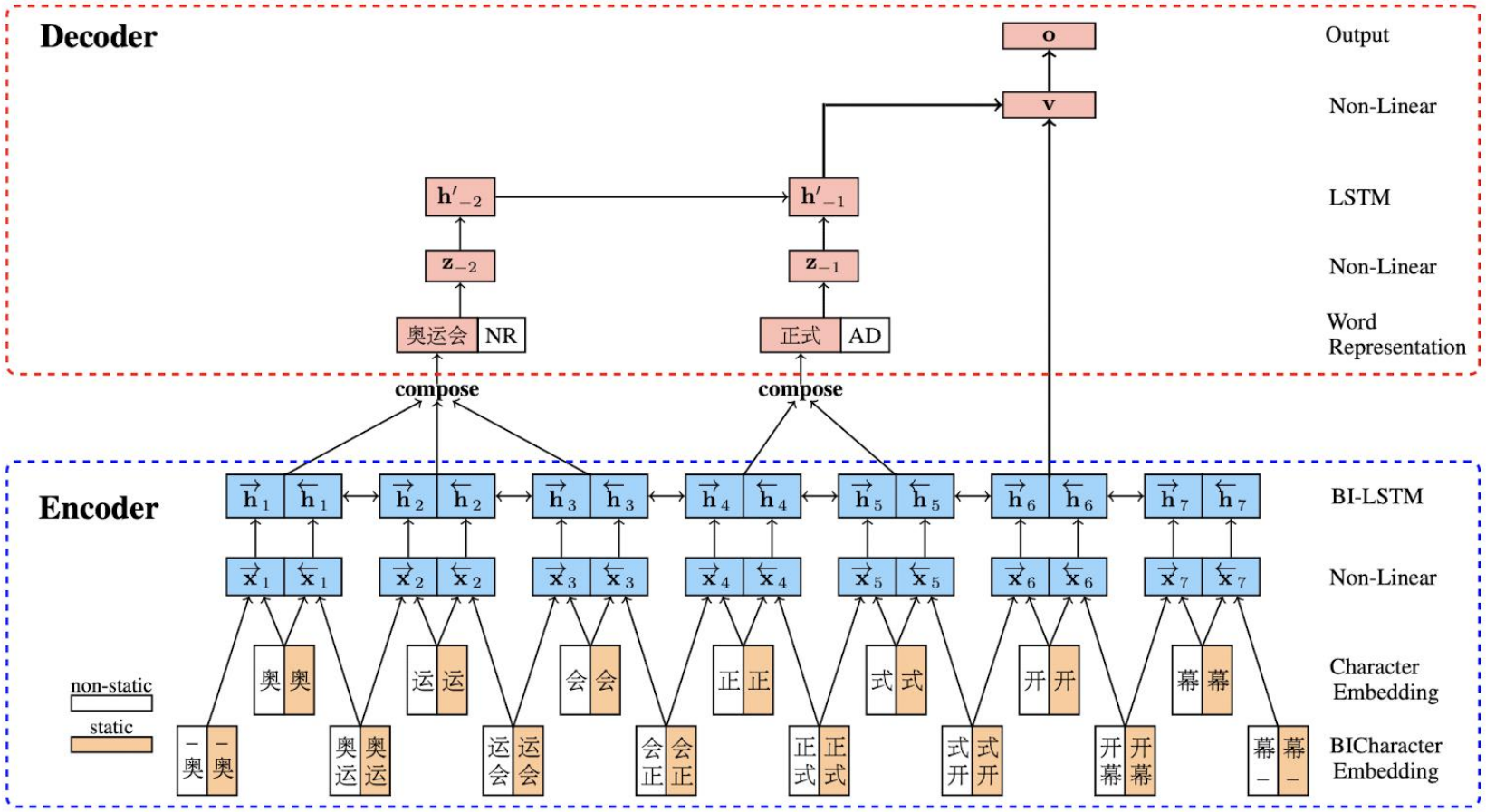
Transition-based Models[6]

Step	Action	State	
		stack($\cdots w_{-2}\|t_{-2}\ \ w_{-1}\|t_{-1}$)	queue($c_0 c_1 \cdots$)
0	-	ϕ	ao yun $\cdots$
1	SEP(NR)	奥(ao)|NR	运(yun) 会(hui) $\cdots$
2	APP	奥运(ao yun)|NR	会(hui) 正(zheng) $\cdots$
3	APP	奥运会(ao yun hui)|NR	正(zheng) 式(shi) $\cdots$
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

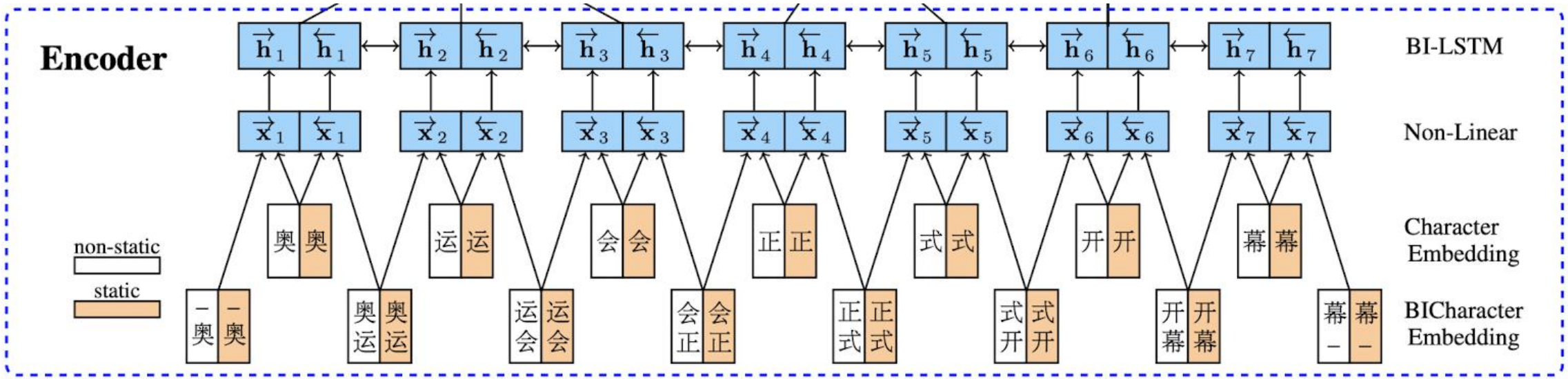
Table I

- SEP(t)：将队列的第一个字符移动到堆栈为带有 POS 标记 t 的新（子）词。
- APP：将队列的第一个字符移入堆栈，将其附加到栈顶（子）字。

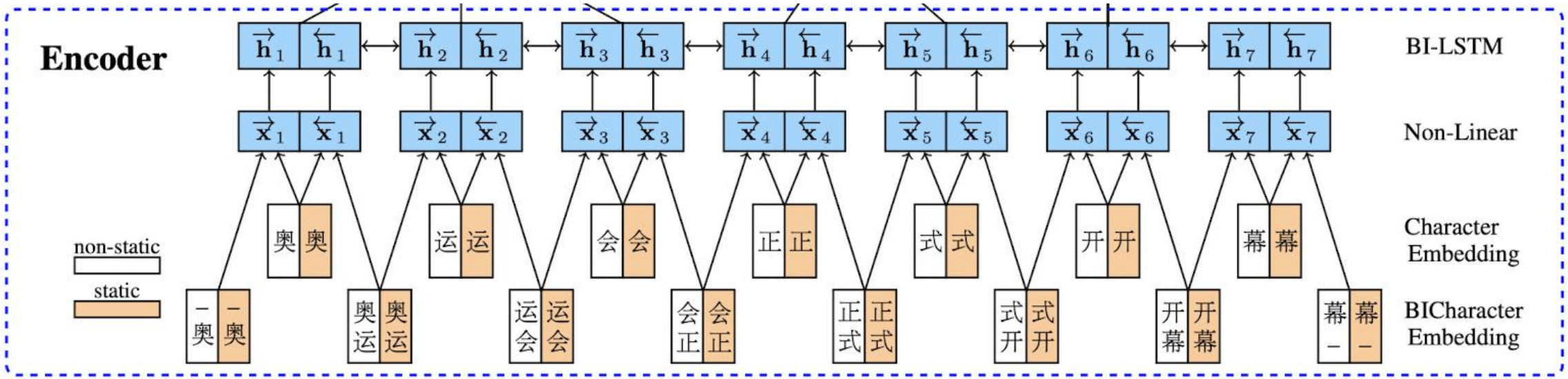
Transition-based Models[6]



Transition-based Models[6]



Transition-based Models[6]



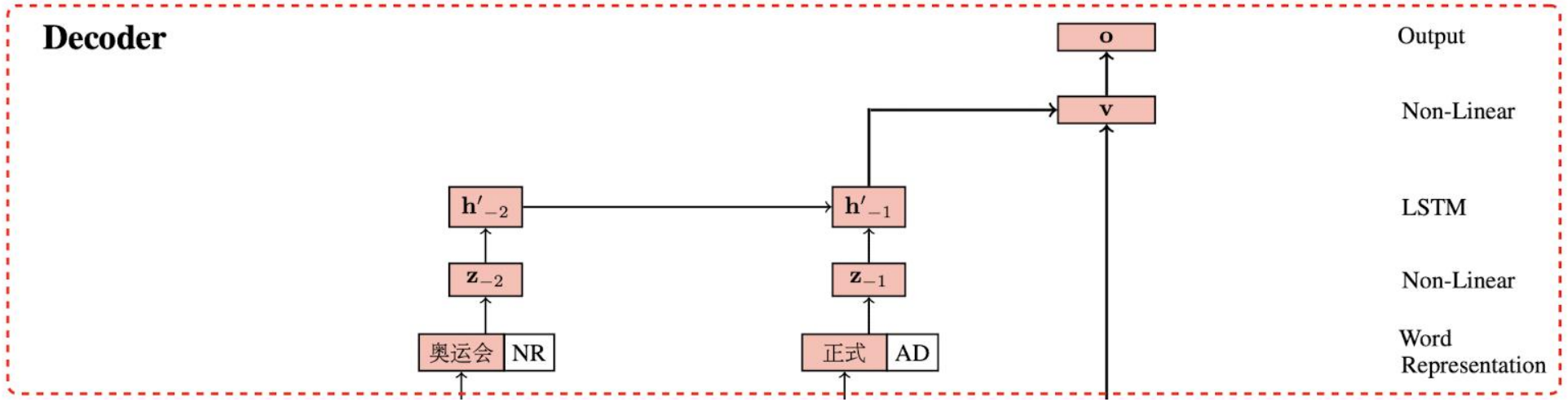
$$\overrightarrow{\mathbf{x}}_t = \tanh(\mathbf{W}_c[\mathbf{E}^c_{c_t} \oplus \tilde{\mathbf{E}}^c_{c_t}, \mathbf{E}^{bc}_{c_{t-1}c_t} \oplus \tilde{\mathbf{E}}^{bc}_{c_{t-1}c_t}] + \mathbf{b}_c)$$

$$\overleftarrow{\mathbf{x}}_t = \tanh(\mathbf{W}_c[\mathbf{E}^c_{c_t} \oplus \tilde{\mathbf{E}}^c_{c_t}, \mathbf{E}^{bc}_{c_t c_{t+1}} \oplus \tilde{\mathbf{E}}^{bc}_{c_t c_{t+1}}] + \mathbf{b}_c)$$

Look-up tables

- Pre-train static
- Learnable non-static

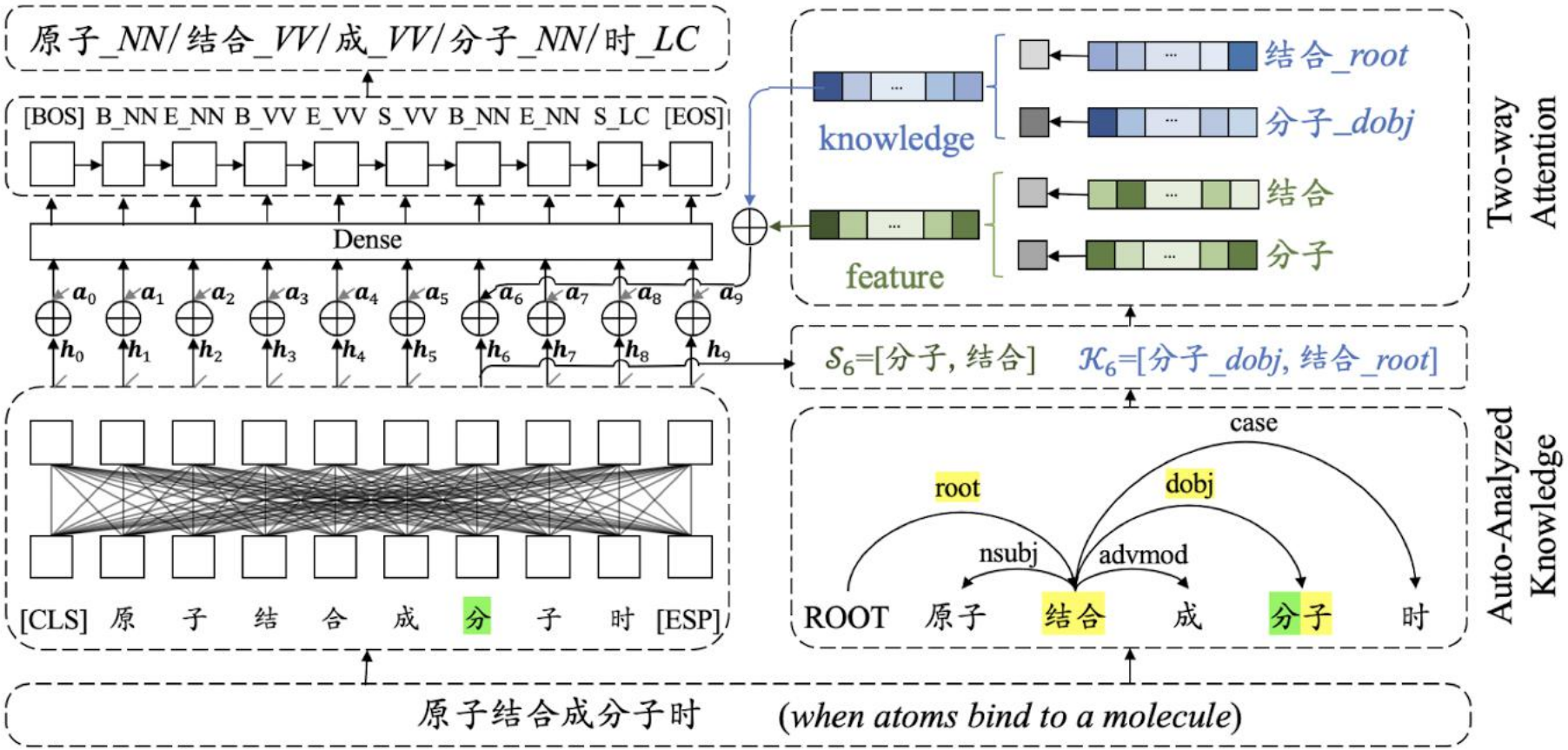
Transition-based Models[6]



词的表示 Z_{-i}

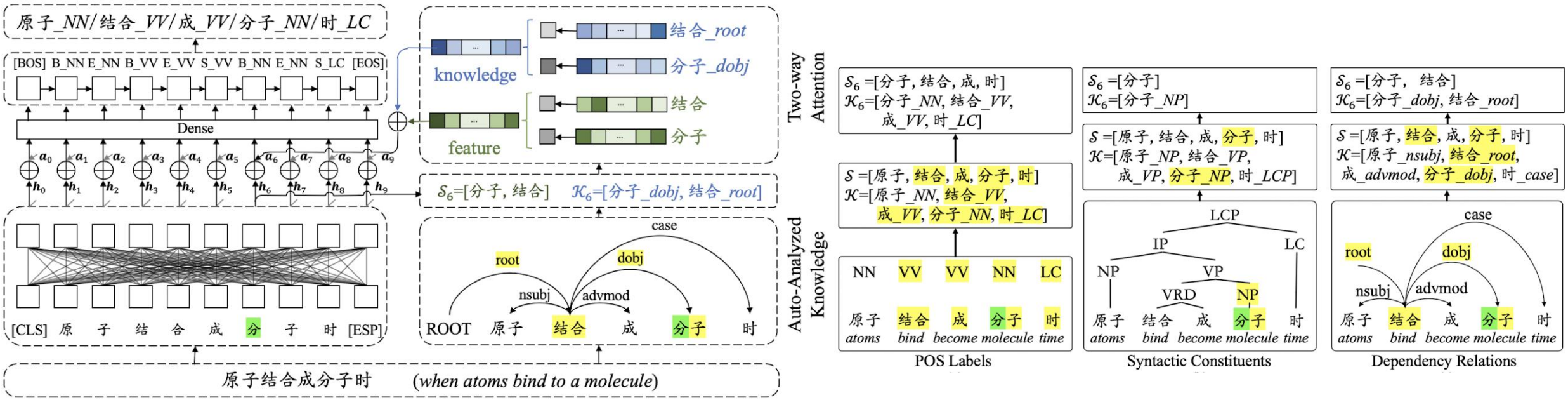
- 词性信息
- 词的字符序列

```
def forward(self, x, train=False):
    """
    :param x:
    :param train:
    :return:
    """
    encoder = self.encoder(x)
    decoder_out, state = self.decoder(x, encoder, train=train)
    return decoder_out, state
```



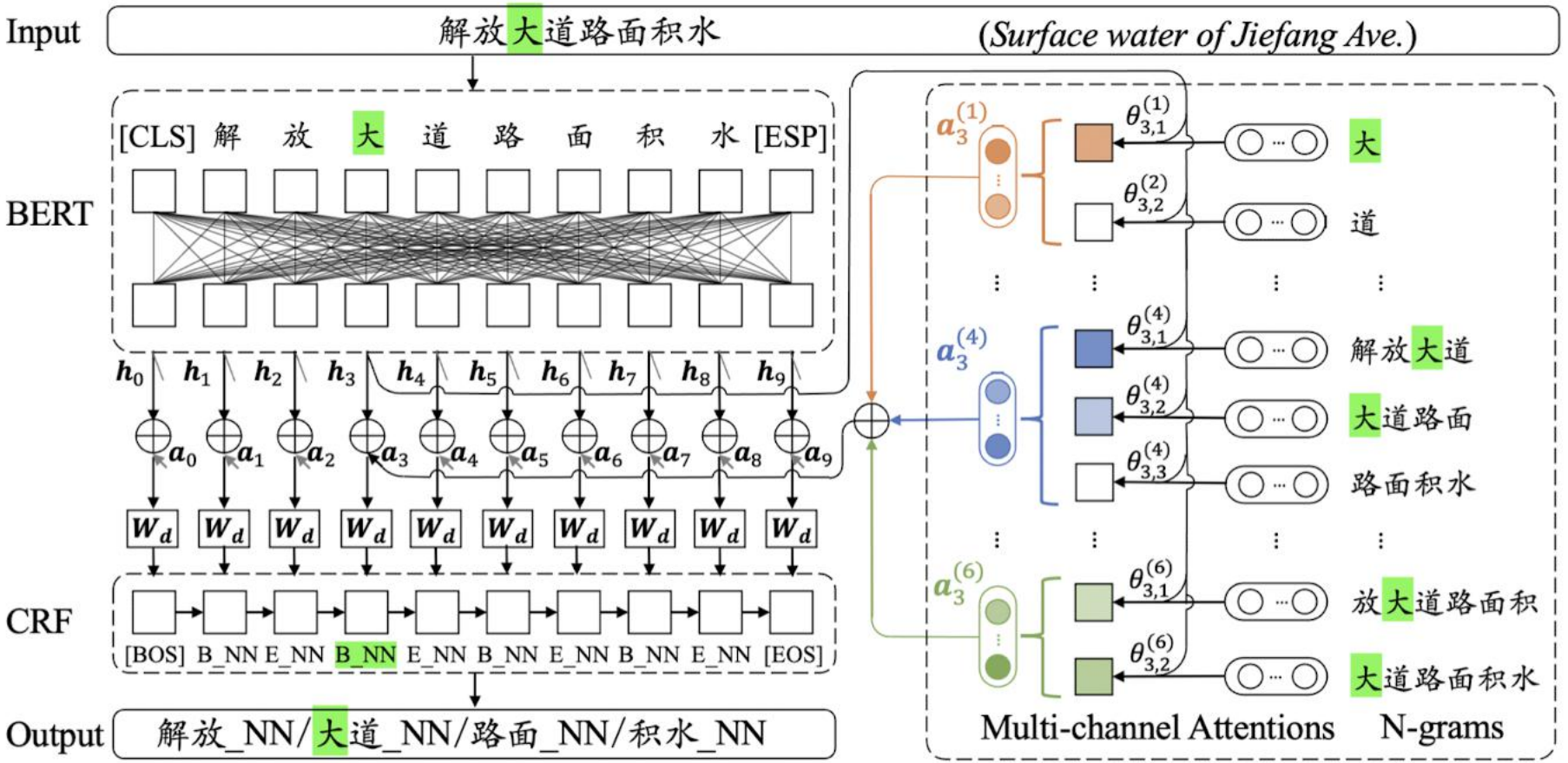
TwASP: Bert-Based[7] 2020SOTA

对自动获取的上下文特征和句法知识通过two-way attention mechanism对特征进行处理，预测每个字的分词和词性标签，



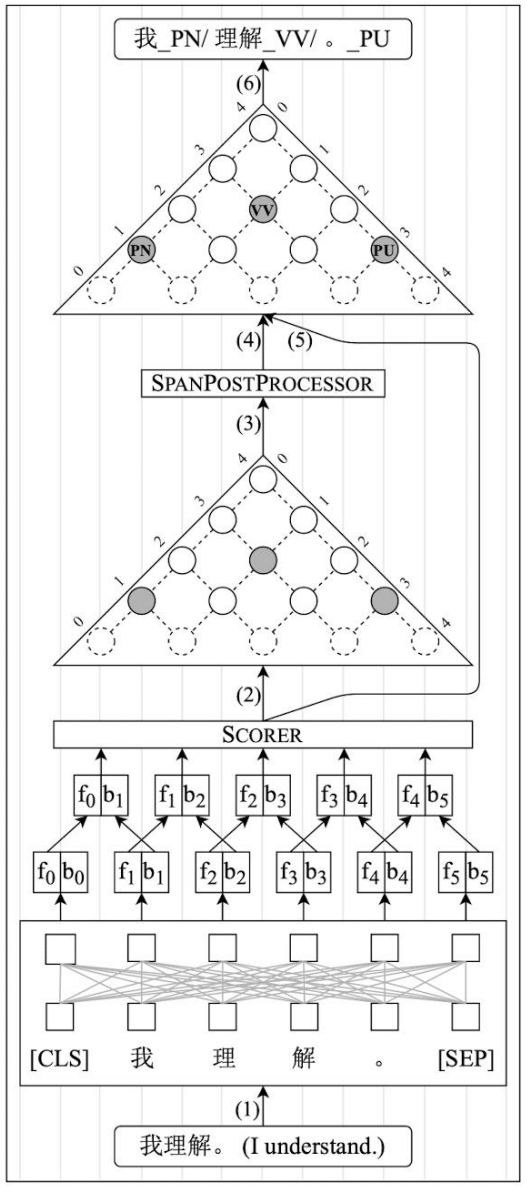
TwASP: Bert-Based[7]

对自动获取的上下文特征和句法知识通过two-way attention mechanism对特征进行处理，预测每个字的分词和词性标签，



McATT[8]:

通过计算字符在不同通道下的attention对该字在模型中的特征进行处理



	CTB5		CTB6		CTB7		CTB9		UD1		UD2	
	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag
Jiang et al. (2008)	97.85	93.41	-	-	-	-	-	-	-	-	-	-
Kruengkrai et al. (2009)	97.87	93.67	-	-	-	-	-	-	-	-	-	-
Sun (2011)	98.17	94.02	-	-	-	-	-	-	-	-	-	-
Wang et al. (2011)	98.11	94.18	95.79	91.12	95.65	90.46	-	-	-	-	-	-
Shen et al. (2014)	98.03	93.80	-	-	-	-	-	-	-	-	-	-
Kurita et al. (2017)	98.41	94.84	-	-	96.23	91.25	-	-	-	-	-	-
Shao et al. (2017)	98.02	94.38	-	-	-	-	96.67	92.34	95.16	89.75	95.09	89.42
Zhang et al. (2018)	98.50	94.95	96.36	92.51	96.25	91.87	-	-	-	-	-	-
Tian et al. (2020a) (BERT)	98.77	96.77	97.39	94.99	**97.32**	94.28	97.75	94.87	98.32	95.60	98.33	95.46
Tian et al. (2020a) (ZEN)	**98.81**	**96.92**	97.47	95.02	97.31	94.32	97.77	94.88	**98.33**	**95.69**	**98.18**	95.49
SPANSEGTAG (BERT)	98.67	96.77	**97.53**	**95.04**	97.30	**94.50**‡	**97.86**	**95.22**‡	98.06	95.59	98.12	**95.54**

SPANSEGTG[9]:

Bert+ two stage Span Labeling

Future Trends

大规模预训练模型?

预训练任务?

迁移学习?

集成算法?

Reference

[1]Ng H T, Low J K. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based?[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 277-284.
[2]Jiang W, Huang L, Liu Q, et al. A cascaded linear model for joint chinese word segmentation and part-of-speech tagging[C]//Proceedings of ACL-08: HLT. 2008: 897-904.
[3]Sun W. A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 1385-1394.
[4]Shao Y, Hardmeier C, Tiedemann J, et al. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF[J]. arXiv preprint arXiv:1704.01314, 2017.[5]A Simple and Effective Neural Model for Joint Word Segmentation and POS Tagging
[5]Kurita S, Kawahara D, Kurohashi S. Neural joint model for transition-based Chinese syntactic analysis[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1204-1214.
[6]Zhang M, Yu N, Fu G. A simple and effective neural model for joint word segmentation and POS tagging[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(9): 1528-1538.
[7]Tian Y, Song Y, Ao X, et al. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8286-8296.
[8]Tian Y, Song Y, Xia F. Joint Chinese word segmentation and part-of-speech tagging via multi-channel attention of character n-grams[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 2073-2084.
[9]Nguyen D V, Vo L B, Tran N L, et al. Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-stage Span Labeling[J]. arXiv preprint arXiv:2112.09488, 2021.

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

4

常见的数据集与技术平台

演讲：蒲沅东

Datasets

Datasets

1、SIGHAN Bakeoff 2005 MSRA,560KB http://sighan.cs.uchicago.edu/bakeoff2005/

2、SIGHAN Bakeoff 2005 PKU, 510KB http://sighan.cs.uchicago.edu/bakeoff2005/

3、SIGHAN Bakeoff 2005 CityU, 510KB http://sighan.cs.uchicago.edu/bakeoff2005/

4、SIGHAN Bakeoff 2005 AS, 510KB http://sighan.cs.uchicago.edu/bakeoff2005/

5、人民日报 2014, 65MB https://pan.baidu.com/s/1hq3KKX

6、 MSRA (新闻语料) https://pan.baidu.com/s/1twci0QVBeWXUg06dK47tiA?_at_=1645796271787

7、CTB8 https://link.zhihu.com/?target=https%3A//pan.baidu.com/s/1DCjDOxB0HD2NmP9w1jm8MA

8、weibo https://link.zhihu.com/?target=https%3A//pan.baidu.com/s/1QHoK2ahpZnNmX6X7Y9iCgQ

https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER 也有汇总

Datasets

CTB：简体中文
UD ： 繁体中文

Datasets		Char	Word	Sent	OOV %
CTB5	Train	805K	494K	18K	-
	Dev	12K	7K	350	8.1
	Test	14K	8K	348	3.5
CTB6	Train	1,056K	641K	23K	-
	Dev	100K	60K	2K	5.4
	Test	134K	82K	3K	5.6
CTB7	Train	1,160K	718K	31K	-
	Dev	387K	237K	10K	5.5
	Test	399K	245K	10K	5.2
CTB9 (general)	Train	2,643K	1,696K	106K	-
	Dev	210K	136K	10K	2.9
	Test	379K	242K	16K	3.1
UD	Train	156K	99K	4K	-
	Dev	20K	13K	500	12.1
	Test	19K	12K	500	12.4
CTB9 (genres)	BC	275K	184K	12K	2.8
	BN	483K	287K	10K	5.1
	CS	228K	160K	17K	5.5
	DF	644K	421K	20K	3.7
	MZ	403K	258K	8K	7.5
	NW	427K	251K	10K	5.1
	SC	430K	304K	44K	4.0
	WB	342K	210K	10K	5.3

Datasets

名称	规模	创建日期	单位
MSR	2368391词，4050469字	2005年	微软亚洲研究院
PKU	1109947词，1826448字	2005年	北京大学
AS	5449698词，8368050字	2005年	台湾中央研究院
CityU	1455629词，2403355字	2005年	香港城市大学

SIGHAN

The Second International Chinese Word Segmentation Bakeoff

Corpora from the following organizations were used:

- CKIP, Academia Sinica, Taiwan
- City University of Hong Kong, Hong Kong SAR
- Beijing Universty, China
- Microsoft Research, China

HanLP: Han Language Processing

面向生产环境的多语种自然语言处理工具包

- 功能完善
- 性能高效
- 语料时新
- 架构清晰

https://github.com/hankcs/HanLP

THULAC：一个高效的中文词法分析工具包

Punctuation as Implicit Annotations for Chinese Word Segmentation

- 能力强
- 准确率高
- 速度快

msr_test（560KB）

Algorithm	Time	Precision	Recall
LTP-3.2.0	3.21s	0.867	0.896
ICTCLAS(2015版)	0.55s	0.869	0.914
jieba	0.26s	0.814	0.809
THULAC	0.62s	0.877	0.899

pku_test（510KB）

Algorithm	Time	Precision	Recall
LTP-3.2.0	3.83s	0.960	0.947
ICTCLAS(2015版)	0.53s	0.939	0.944
jieba	0.23s	0.850	0.784
THULAC	0.51s	0.944	0.908

https://github.com/thunlp/THULAC

pkuseg：一个多领域中文分词工具包

- 多领域分词
- 更高的分词准确率
- 支持用户自训练模型
- 支持词性标注。

Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection

Xu Sun[†], **Houfeng Wang**[‡], **Wenjie Li**[†]
[†]Department of Computing, The Hong Kong Polytechnic University
[‡]Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, China
{csxsun, cswjli}@comp.polyu.edu.hk wanghf@pku.edu.cn

CRF+ADF

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

pkuseg：一个多领域中文分词工具包

- 多领域分词
- 更高的分词准确率
- 支持用户自训练模型
- 支持词性标注。

Domain	Vocabulary Size
Medicine	447K
Location	117K
Name	105K
Idiom	50K
Organization	31K
Training Words	100K
total	850K

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

pkuseg：一个多领域中文分词工具包

- 多领域分词
- 更高的分词准确率
- 支持用户自训练模型
- 支持词性标注。

MSRA	Precision	Recall	F-score
jieba	87.01	89.88	88.42
THULAC	95.60	95.91	95.71
pkuseg	96.94	96.81	**96.88**

WEIBO	Precision	Recall	F-score
jieba	87.79	87.54	87.66
THULAC	93.40	92.40	92.87
pkuseg	93.78	94.65	**94.21**

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

NLPIR汉语分词系统

中英文混合分词功能

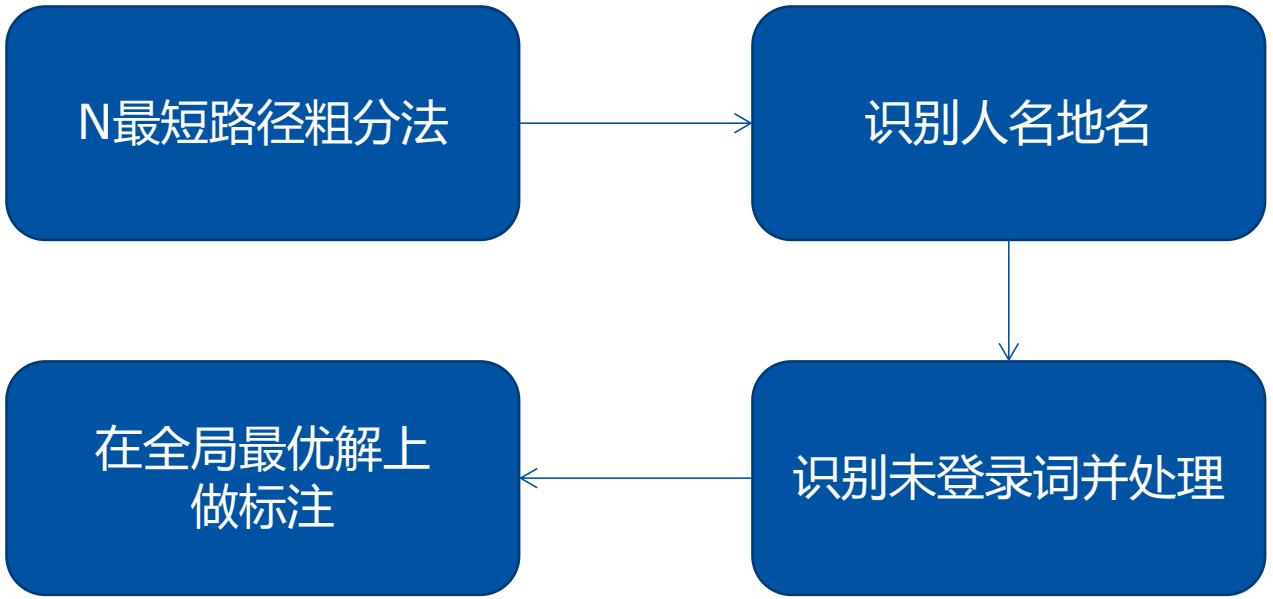
关键词提取功能

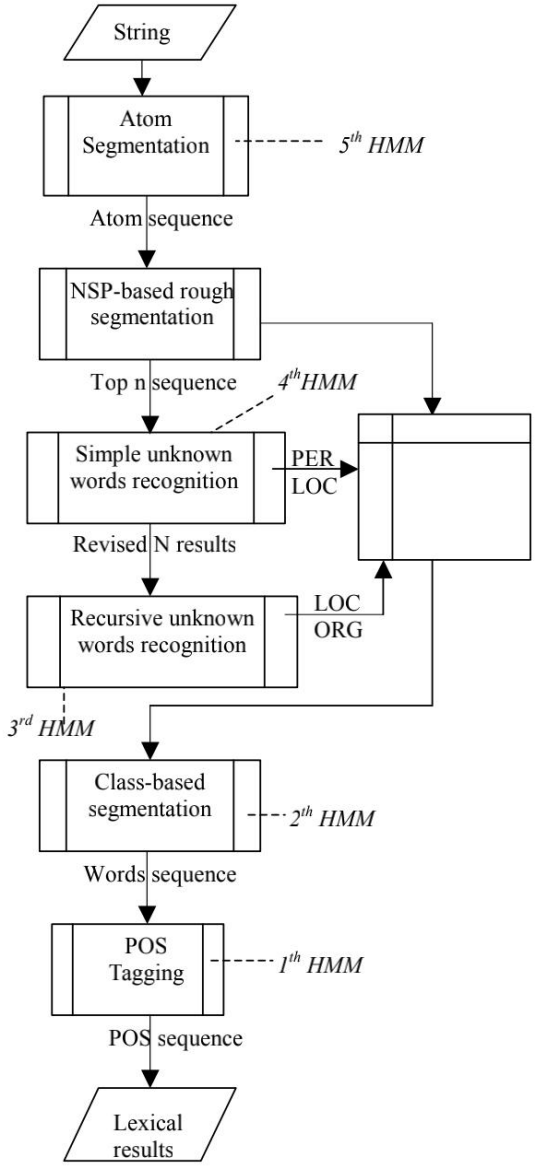
新词识别与自适应分词功能

用户专业词典功能

http://kgb.lingjoin.com/nlpir/

层叠隐马尔可夫模型


N最短路径粗分法
识别人名地名
识别未登录词并处理
在全局最优解上做标注


String
Atom Segmentation
5th HMM
Atom sequence
NSP-based rough segmentation
Top n sequence
4th HMM
Simple unknown words recognition
PER
LOC
Revised N results
Recursive unknown words recognition
LOC
ORG
3rd HMM
Class-based segmentation
2th HMM
Words sequence
POS Tagging
1th HMM
POS sequence
Lexical results

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Demo

演讲：李翰东

1.中文分词

2.词性标注

3.联合模型

4.实战

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

数据集（SIGHAN2005）

名称	规模	创建日期	单位
MSR	2368391词，4050469字	2005年	微软亚洲研究院
PKU	1109947词，1826448字	2005年	北京大学
AS	5449698词，8368050字	2005年	台湾中央研究院
CityU	1455629词，2403355字	2005年	香港城市大学

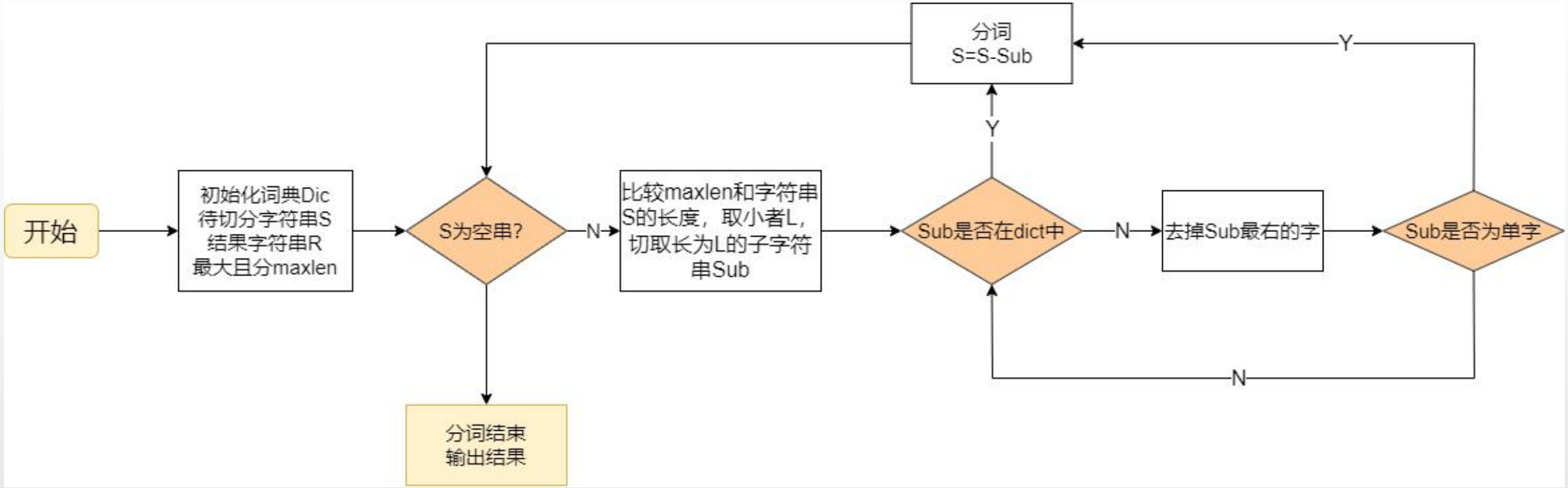
SIGHAN

The Second International Chinese Word Segmentation Bakeoff

Corpora from the following organizations were used:

- CKIP, Academia Sinica, Taiwan
- City University of Hong Kong, Hong Kong SAR
- Beijing Universty, China
- Microsoft Research, China

基于词典


开始
初始化词典Dic
待切分字符串S
结果字符串R
最大且分maxlen
S为空串?
N
比较maxlen和字符串
S的长度，取小者L，
切取长为L的子字符
串Sub
Sub是否在dict中
N
去掉Sub最右的字
Sub是否为单字
Y
分词
S=S-Sub
Y
N
分词结束
输出结果

基于词典 demo

```
def mm_tokenize(self, text):
    tokens = []                 # 定义一个空列表来保存切分的结果
    index = 0                   # 切分
    text_length = len(text)
    windows_size = min(self.window_size, text_length)

    while text_length > index:     # 循环结束判定条件
        for size in range(windows_size + index, index, -1):    # 根据窗口大小循环，直到找到符合的进行下一次循环
            piece = text[index: size]       # 被匹配字段
            if piece in self.dic:     # 如果需要被匹配的字段在词典中的话匹配成功，新的index为新的匹配字段的起始位置
                index = size - 1
                break
        index += 1
        tokens.append(piece)          # 将匹配到的字段保存起来

    return tokens
```

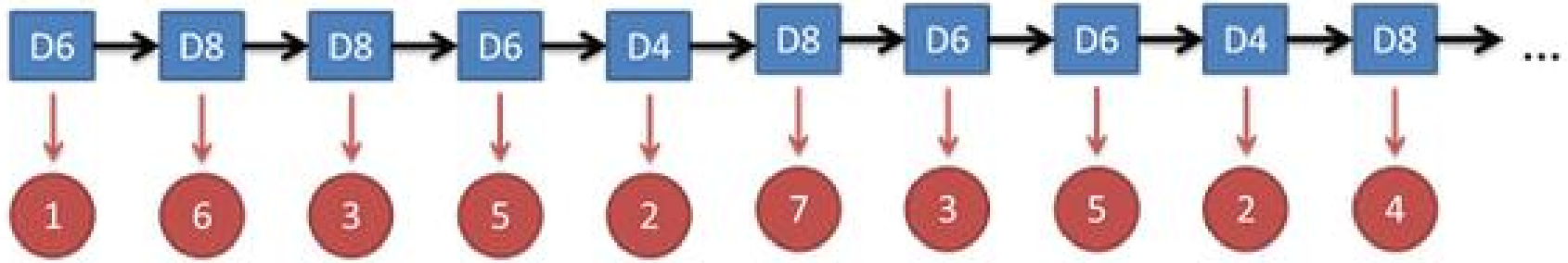
基于词典 demo

```
lihan on D:/python code/nlp
# python .\BiDirectionMatching.py
text:
研究生研究生命的起源
max_len:
4
正向最大匹配分词结果: ['研究生', '研究生', '命', '的', '起源']
逆向最大匹配分词结果: ['研究生', '研究', '生命', '的', '起源']
双向最大匹配得到的结果:  ['研究生', '研究', '生命', '的', '起源']
```

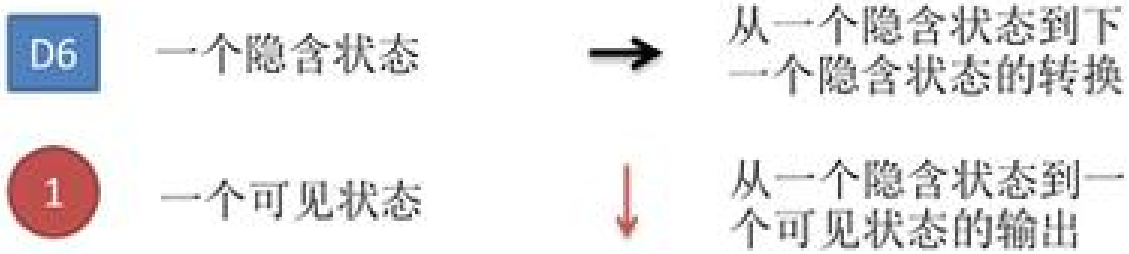
```
lihan on D:/python code/nlp
# python .\BiDirectionMatching.py
text:
欢迎新老师生前来就餐
max_len:
4
正向最大匹配分词结果: ['欢迎', '新', '老师', '生前', '来', '就餐']
逆向最大匹配分词结果: ['欢', '迎新', '老', '师生', '前来', '就餐']
双向最大匹配得到的结果:  ['欢迎', '新', '老师', '生前', '来', '就餐']
```

基于统计（HMM）

隐马尔可夫模型示意图



图例说明：



```
{'B': 0.03942387059367564, 'E': 0.0, 'M': 0.0, 'S': 0.017235288687604527}
> special variables
> function variables
  'B': 0.03942387059367564
  'E': 0.0
  'M': 0.0
  'S': 0.017235288687604527
  len(): 4

按住 Alt 键可切换到编辑器语言悬停
```

```
{'B': {'B': 0.0, 'E': 0.8364130260541732, 'M': 0.16358697394582675, 'S': 0.0}, 'E': {'B…
> special variables
> function variables
> 'B': {'B': 0.0, 'E': 0.8364130260541732, 'M': 0.16358697394582675, 'S':…
> 'E': {'B': 0.48704973574535404, 'E': 0.0, 'M': 0.0, 'S': 0.510596497610…
> 'M': {'B': 0.0, 'E': 0.0, 'M': 0.46077901049260434, 'S': 0.539220989507…
> 'S': {'B': 0.5250087282625008, 'E': 0.0, 'M': 0.0, 'S': 0.4222096505446…
  len(): 4

按住 Alt 键可切换到编辑器语言悬停
```

```
{'B': {'充': 0.0008699758352121225, '希': 0.0009286625249078 81, '世': 0.002991934383933_
> special variables
> function variables
> 'B': {'充': 0.0008699758352121225, '希': 0.0009286625249078 81, '世': 0.00
> 'E': {'向': 0.0013851036391454197, '满': 0.0004788092786351 6617, '望': 0.
> 'M': {'九': 0.0019917678372482 08, '八': 0.0014759142966659382, '共': 0.00
> 'S': {'的': 0.093931059876654 3, '新': 0.0031308290482764452, '年': 0.0067
  len(): 4
```

```
from segment.hmm import *

segger = HMMSegger()
segger.load_data("./data/train1.utf8")
# print(hmm.data.readline())
segger.train()
print(segger.cut("2001年新年钟声即将敲响。人类社会前进的航船就要驶入21世纪的新航程。"))
```

```
Microsoft Windows [版本 10.0.22000.527]
(c) Microsoft Corporation。保留所有权利。

D:\python code\demo>python -u "d:\python code\demo\HMM\train.py"
['2001', '年', '新年', '钟', '声即', '将', '敲响', '。', '人类', '社会', '前进', '的', '航船', '就', '
要驶', '入', '21世纪', '的', '新', '航程', '。']

D:\python code\demo>C:/Users/lihan/anaconda3/Scripts/activate
```

基于统计（N-Gram）

当 n=1, 一个一元模型（unigram model)即为：

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i)$$

当 n=2, 一个二元模型（bigram model)即为：

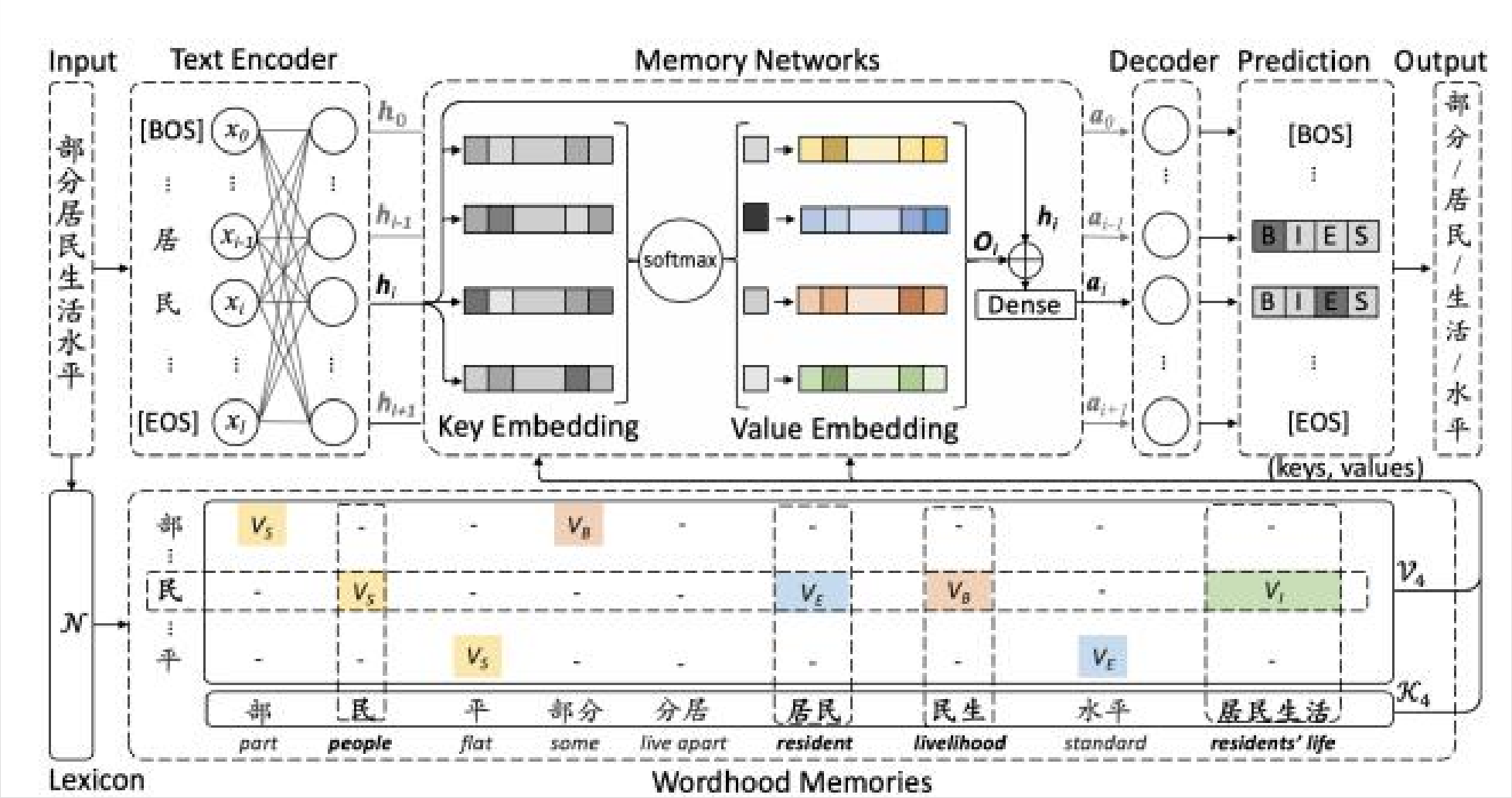
$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-1})$$

当 n=3, 一个三元模型（trigram model)即为

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-2}w_{i-1})$$

```python
def word_segment_ngram(self, input_str):
    segments = self.word_break(input_str, self.corpus.token_dict.keys())
    best_segment = []
    best_score = math.inf
    for seg in segments:
        score = 0
        for word in seg:
            if word in self.corpus.token_dict.keys():
                score += self.corpus.token_dict[word]
            else:
                score += round(-math.log(1e-10), 1)
        if score < best_score:
            best_score = score
            best_segment = seg
    if len(best_segment)==0:
        return [input_str]
    return best_segment
```

神经网络WMSeg（Bert/Zen + CRF）



神经网络WMSeg（Bert/Zen + CRF）

```
03/09/2022 10:05:22 - INFO - pytorch_pretrained_zen.modeling -   loading weight
03/09/2022 10:05:22 - INFO - pytorch_pretrained_zen.modeling -   loading confi
03/09/2022 10:05:22 - INFO - pytorch_pretrained_zen.modeling -   Model config
  "attention_probs_dropout_prob": 0.1,
  "directionality": "bidi",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "num_hidden_word_layers": 6,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "type_vocab_size": 2,
  "vocab_size": 21128,
  "word_size": 104089
}

03/09/2022 10:05:42 - INFO - __main__ -   # of trainable parameters: 354168640
03/09/2022 10:05:42 - INFO - __main__ -   ***** Running training *****
03/09/2022 10:05:42 - INFO - __main__ -     Num examples = 30769
```

```
19980101-01-001-001/m  迈向/v  充满/v  希望/n  的/u  新/a  世纪/n  —/w  一九九八年/t  新年/t  讲话/n
19980101-01-001-002/m  中共中央/nt  总书记/n  、/w  国家/n  主席/n  江/nr  泽民/nr
19980101-01-001-003/m  （/w  一九九七年/t  十二月/t  三十一日/t  ）/w
19980101-01-001-004/m  １２月/t  ３１日/t  ，/w  中共中央/nt  总书记/n  、/w  国家/n  主席/n  江/nr  泽
19980101-01-001-005/m  同胞/n  们/k  、/w  朋友/n  们/k  、/w  女士/n  们/k  、/w  先生/n  们/k  ：/w
19980101-01-001-006/m  在/p  １９９８年/t  来临/v  之际/f  ，/w  我/r  十分/m  高兴/a  地/u  通过/p  [
19980101-01-001-007/m  １９９７年/t  ，/w  是/v  中国/ns  发展/vn  历史/n  上/f  非常/d  重要/a  的/u
19980101-01-001-008/m  在/p  这/r  一/m  年/q  中/f  ，/w  中国/ns  的/u  改革/vn  开放/vn  和/c  现代
19980101-01-001-009/m  在/p  这/r  一/m  年/q  中/f  ，/w  中国/ns  的/u  外交/n  工作/vn  取得/v  了/
19980101-01-001-010/m  １９９８年/t  ，/w  中国/ns  人民/n  将/d  满怀信心/l  地/u  开创/v  新/a  的/u
19980101-01-001-011/m  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
19980101-01-001-012/m  台湾/ns  是/v  中国/ns  领土/n  不可分割/l  的/u  一/m  部分/n  。/w  完成/v  祖
19980101-01-001-013/m  环顾/v  全球/n  ，/w  日益/d  密切/a  的/u  世界/n  经济/n  联系/vn  ，/w  日新
19980101-01-001-014/m  [中国/ns  政府/n]nt  将/d  继续/v  坚持/v  奉行/v  独立自主/i  的/u  和平/n  外
19980101-01-001-015/m  在/p  这/r  辞旧迎新/l  的/u  美好/a  时刻/n  ，/w  我/r  祝/v  大家/r  新年/t
19980101-01-001-016/m  谢谢/v  ！/w  （/w  新华社/nt  北京/ns  １２月/t  ３１日/t  电/n  ）/w

19980101-01-002-001/m  在/p  十五大/j  精神/n  指引/vn  下/f  胜利/vd  前进/v  —/w  元旦/t  献辞/n
19980101-01-002-002/m  我们/r  即将/d  以/p  丰收/vn  的/u  喜悦/an  送/v  走/v  牛年/t  ，/w  以/p  昂
19980101-01-002-003/m  刚刚/d  过去/v  的/u  一/m  年/q  ，/w  大气磅礴/i  ，/w  波澜壮阔/i  。/w  在/
19980101-01-002-004/m  １９９８年/t  ，/w  是/v  全面/ad  贯彻/v  落实/v  党/n  的/u  十五大/j  提出/v
19980101-01-002-005/m  今年/t  是/v  党/n  的/u  十一/m  届/q  三中全会/j  召开/v  ２０/m  周年/q  ，/
```

PFR语料库：
人民日报1998年语料

基于统计（HMM）

```
from lable import *


l = Lable()
l.train()
text = "研究生研究生命的起源"
ans = l.do_lable(text)
print(ans)
```

```
(base) ➜  HMMlable python main.py
研究生/n 研究/vn 生命/n 的/u 起源/n
(base) ➜  HMMlable
```

神经网络McASP （Bert/Zen + CRF）

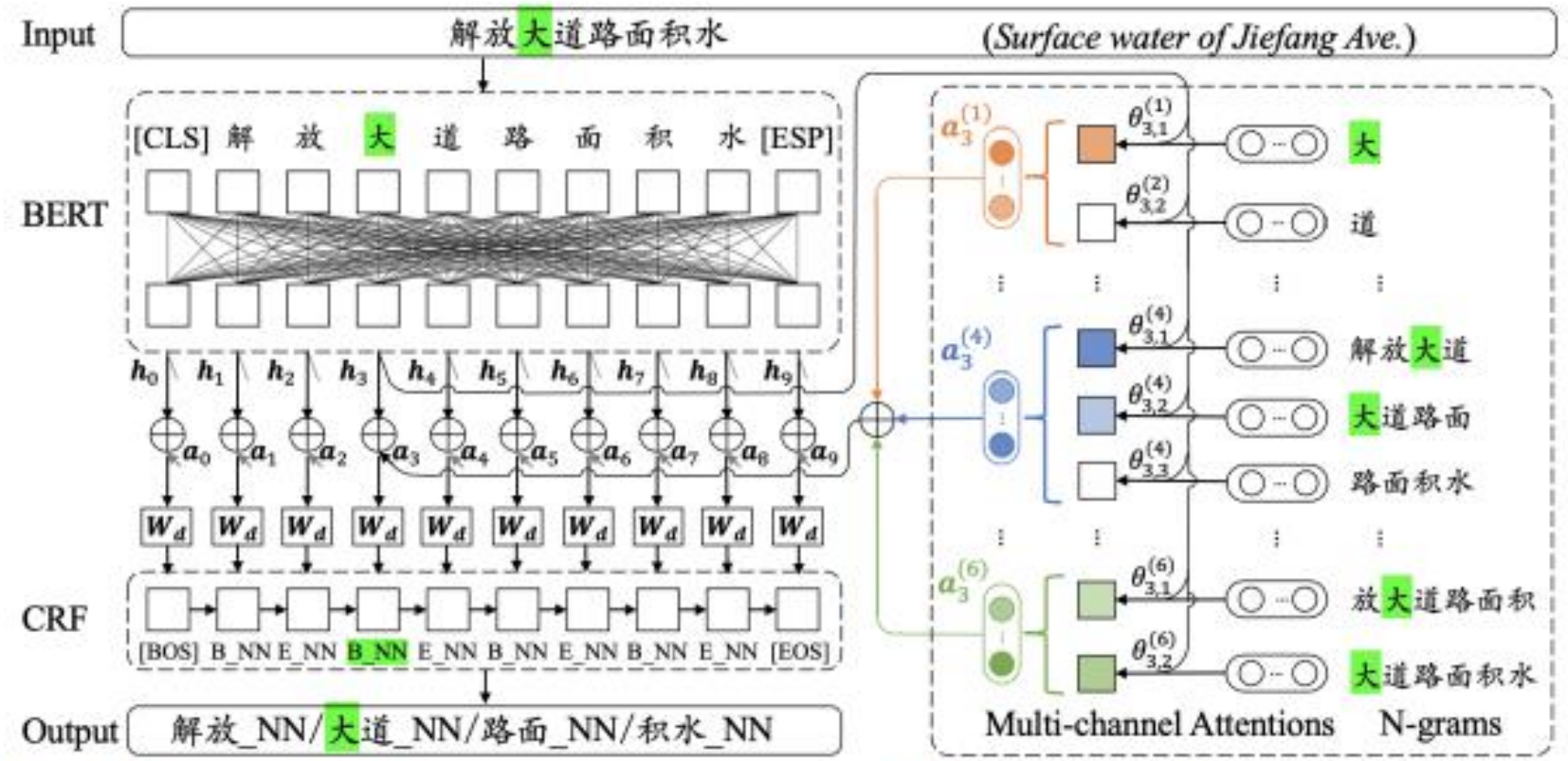


Figure 2: The overall architecture of our character-based model for the joint CWS and POS tagging with an example input and output. On the left is the backbone model following the sequence labeling paradigm; on the right is the multi-channel attention module with n-grams categorized by their length. Different attention channels for n-grams associated with "大" (*big*) are highlighted with distinct colors.

神经网络McASP （Bert/Zen + CRF）

共同创造美好的新世纪—二○○一年新年贺词

（二○○○年十二月三十一日）（附图片1张）

女士们，先生们，同志们，朋友们：

2001年新年钟声即将敲响。人类社会前进的航船就要驶入21世纪的新航程。中国人民进入了向现代化建设第三步战略目标迈进的新征程。

神经网络McASP（Bert/Zen + CRF）

共同_RB 创造_VV 美好_JJ 的_DEC 新_PFA 世纪_NN ——_FW 二○○一_CD 年_NNB 新年_NN 贺词_NN
（_ （ 二○○○_CD 年_NNB 十二_CD 月_NNB 三十一_CD 日_NNB ） _）
（_ （ 附_VV 图片_NN 1_CD 张_NNB ） _）
女士_NN 们_SFN ，_EC 先生_NN 们_SFN ，_EC 同志_NN 们_SFN ，_EC 朋友_NN 们_SFN ：_:
2001_CD 年_NNB 新年_NN 钟声_NN 即将_RB 敲响_VV 。_. 人类_NN 社会_NN 前进_VV 的_DEC 航船_NN 就_RB 要_MD 驶入_VV 21_CD 世纪_NNB 的_DEC 新_PFA 航程_NN 。_. 中国_NNP 人民_NN 进入_VV 了_AS 向_IN 现代_NN 化_SFN 建设_NN 第三_CD 步_NNB 战略_NN 目标_NN 迈进_VV 的_DEC 新_PFA 征程_NN 。_.

method	Recall	Precision	F	R_oov	R_iv	time
MM	0.854	0.832	0.843	0.847	0.942	0.37
RMM	0.854	0.832	0.843	0.847	0.937	**0.36**
BiMM	0.855	0.834	0.845	0.849	0.942	0.76
N-Gram	0.772	0.884	0.824	0.764	0.872	92.24
HMM	0.782	0.753	0.767	0.772	0.919	214.2
BERT+CRF	**0.932**	**0.969**	**0.950**	**0.930**	**0.959**	17.72

Table 1: pku

method	Recall	Precision	F	R_oov	R_iv	time
MM	0.928	0.907	0.918	0.924	**0.979**	0.47
RMM	0.928	0.907	0.918	0.925	0.97	**0.38**
BiMM	0.93	0.91	0.92	0.927	**0.979**	0.7
N-Gram	0.932	0.921	0.926	0.928	0.975	163.13
HMM	0.78	0.71	0.743	0.766	0.954	250.02
ZEN+CRF	**0.965**	**0.983**	**0.974**	**0.964**	**0.979**	31.24

Table 2: msr

name	Recall	Precision	F	time
jieba	0.768	0.833	0.799	**0.636**
nlpir	**0.937**	**0.935**	**0.936**	0.759
thulac	0.923	0.922	0.923	4.719
ltp	0.841	0.924	0.880	5.526

Table 3: pku

name	Recall	Precision	F	time
jieba	0.810	0.819	0.814	**0.755**
nlpir	**0.912**	**0.866**	**0.888**	0.850
thulac	0.879	0.834	0.856	5.102
ltp	0.842	0.854	0.848	6.015

Table 4: msr

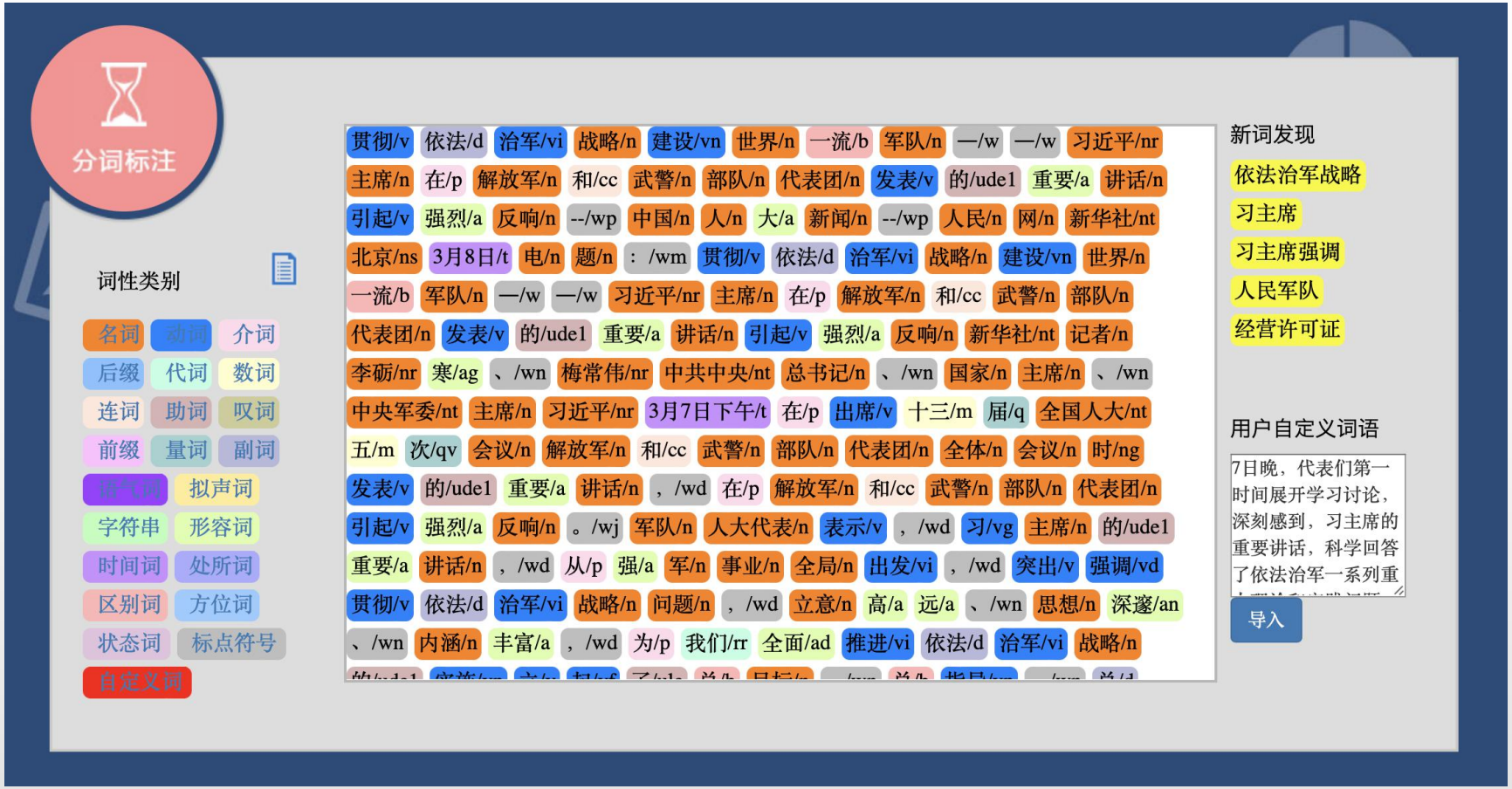
贯彻依法治军战略 建设世界一流军队——习近平主席在解放军和武警部队代表团发表的重要讲话引起强烈反响

2022年03月08日23:34 | 来源：新华网

新华社北京3月8日电 **题：贯彻依法治军战略 建设世界一流军队——习近平主席在解放军和武警部队代表团发表的重要讲话引起强烈反响**

新华社记者李砺寒、梅常伟

中共中央总书记、国家主席、中央军委主席习近平3月7日下午在出席十三届全国人大五次会


分词标注
词性类别
名词 动词 介词
后缀 代词 数词
连词 助词 叹词
前缀 量词 副词
语气词 拟声词
字符串 形容词
时间词 处所词
区别词 方位词
状态词 标点符号
自定义词
贯彻/v 依法/d 治军/vi 战略/n 建设/vn 世界/n 一流/b 军队/n —/w —/w 习近平/nr
主席/n 在/p 解放军/n 和/cc 武警/n 部队/n 代表团/n 发表/v 的/ude1 重要/a 讲话/n
引起/v 强烈/a 反响/n --/wp 中国/n 人/n 大/a 新闻/n --/wp 人民/n 网/n 新华社/nt
北京/ns 3月8日/t 电/n 题/n ：/wm 贯彻/v 依法/d 治军/vi 战略/n 建设/vn 世界/n
一流/b 军队/n —/w —/w 习近平/nr 主席/n 在/p 解放军/n 和/cc 武警/n 部队/n
代表团/n 发表/v 的/ude1 重要/a 讲话/n 引起/v 强烈/a 反响/n 新华社/nt 记者/n
李砺/nr 寒/ag 、/wn 梅常伟/nr 中共中央/nt 总书记/n 、/wn 国家/n 主席/n 、/wn
中央军委/nt 主席/n 习近平/nr 3月7日下午/t 在/p 出席/v 十三/m 届/q 全国人大/nt
五/m 次/qv 会议/n 解放军/n 和/cc 武警/n 部队/n 代表团/n 全体/n 会议/n 时/ng
发表/v 的/ude1 重要/a 讲话/n ，/wd 在/p 解放军/n 和/cc 武警/n 部队/n 代表团/n
引起/v 强烈/a 反响/n 。/wj 军队/n 人大代表/n 表示/v ，/wd 习/vg 主席/n 的/ude1
重要/a 讲话/n ，/wd 从/p 强/a 军/n 事业/n 全局/n 出发/vi ，/wd 突出/v 强调/vd
贯彻/v 依法/d 治军/vi 战略/n 问题/n ，/wd 立意/n 高/a 远/a 、/wn 思想/n 深邃/an
、/wn 内涵/n 丰富/a ，/wd 为/p 我们/rr 全面/ad 推进/vi 依法/d 治军/vi 战略/n
新词发现
依法治军战略
习主席
习主席强调
人民军队
经营许可证
用户自定义词语
7日晚，代表们第一时间展开学习讨论，深刻感到，习主席的重要讲话，科学回答了依法治军一系列重
导入

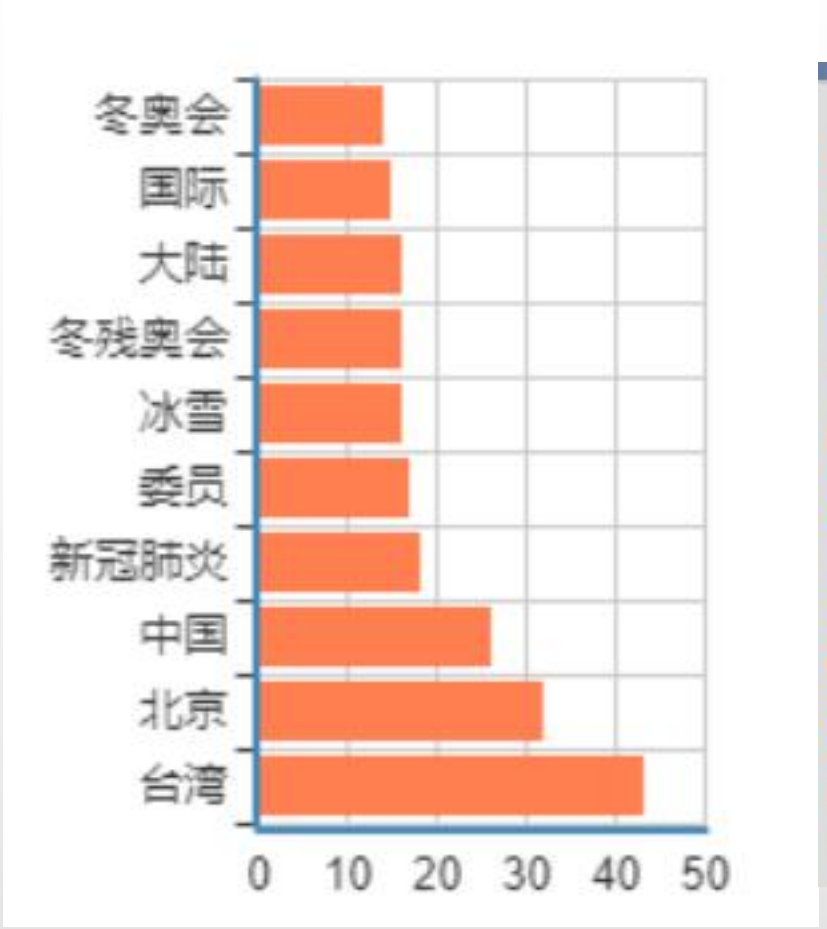
```python
from cProfile import label
from urllib.parse import urljoin
import requests
import bs4
import re

lables = ['要闻','观点']
pages_list =  [['http://tw.people.com.cn/',
                'http://hm.people.com.cn/',
                'http://military.people.com.cn/',
                'http://world.people.com.cn/',],
               ['http://ent.people.com.cn/',
                'http://society.people.com.cn/',
                'http://finance.people.com.cn/',]]
for lable, pages in zip(lables, pages_list):
    f = open("./data/renmin_data/"+lable+".txt", 'w', encoding='utf-8')
    for page in pages:
        r = requests.get(page)
        r.encoding = r.apparent_encoding
        bs = bs4.BeautifulSoup(r.text, "html.parser")
        news = bs.findAll(class_="hdNews")
        for i in news:
            f.write(i.find("a").text)
            f.write(i.find("a").get("href"))
            f.write("\n")
        next_page = bs.find('a', string="下一页")
        if next_page:
            next_page_url = next_page.get("href")
            next_page_url = next_page_url if next_page_url.startswith(
                "http") else urljoin(page, next_page_url)
            r = requests.get(next_page_url)
            r.encoding = r.apparent_encoding
            bs = bs4.BeautifulSoup(r.text, "html.parser")
            news = bs.findAll(class_="hdNews")
            for i in news:
                f.write(i.find("a").text)
                f.write(i.find("a").get("href"))
                f.write("\n")
    f.close()
```

人民日报要闻和观点两个栏目近三个月的所有新闻标题

要闻.txt - 记事本

文件 编辑 查看

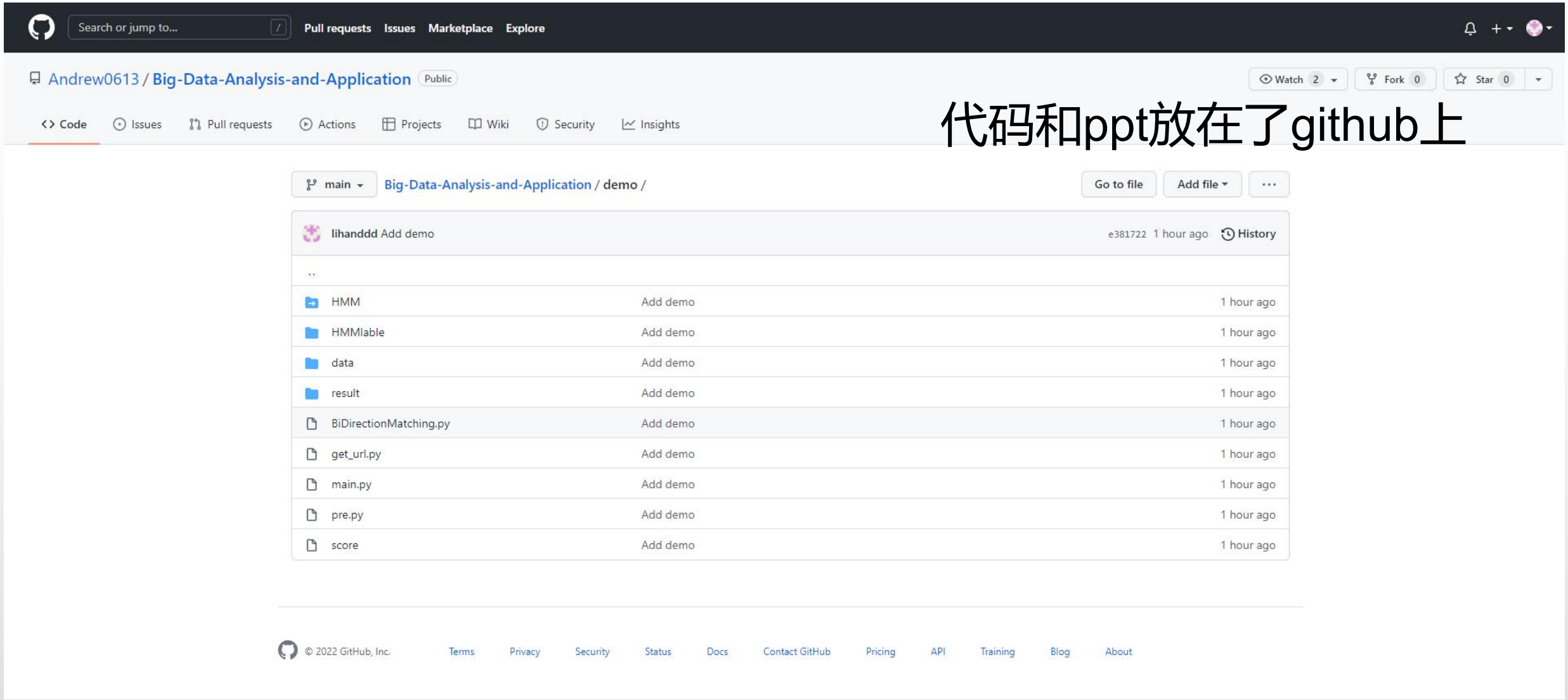
台湾新增71例新冠肺炎确诊病例/n1/2022/0303/c14657-32365370.html
台湾多地无预警停电已造成灾情/n1/2022/0303/c14657-32365365.html
台湾新增49例新冠肺炎确诊病例/n1/2022/0302/c14657-32363744.html
台湾新增44例新冠肺炎确诊病例/n1/2022/0301/c14657-32362756.html
台湾春浪音乐节热闹开场 乐迷期待大陆音乐人回归/n1/2022/0308/c14657-32369454.html
台湾春浪音乐节热闹开场 乐迷期待大陆音乐人回归/n1/2022/0308/c14657-32369275.html
台湾多个团体联合举办春祭缅怀抗日英烈先贤/n1/2022/0304/c14657-32366622.html
大熊猫“圆仔”体重超标 台北市立动物园启动减重计划/n1/2022/0301/c14657-32362137.html
台湾新增56例新冠肺炎确诊病例/n1/2022/0228/c14657-32361791.html
台湾新增68例新冠肺炎确诊病例/n1/2022/0226/c14657-32360325.html
台湾新增68例新冠肺炎确诊个案/n1/2022/0218/c14657-32354797.html
台湾新增52例新冠肺炎确诊病例/n1/2022/0214/c14657-32351456.html
台湾新增83例新冠肺炎确诊病例/n1/2022/0211/c14657-32350069.html
台湾新增54例新冠肺炎确诊病例/n1/2022/0210/c14657-32349173.html
杨毅周：维护台胞权益 助力台胞更好融入国家发展大局/n1/2022/0309/c14657-32370886.html
台盟中央副主席：解决好台青登陆创业就业“水土不服”问题/n1/2022/0308/c14657-32369278.html
江尔雄：我很喜欢《我们同唱一首歌》，欢迎台青来大陆追梦筑梦圆梦/n1/2022/0308/c14657-32369272
两岸同胞富春江畔踏青 携手共绘现代富春山居图/n1/2022/0307/c14657-32368128.html
国台办：“台独”勾结外部势力妄图挑战一个中国原则的图谋必被碾碎/n1/2022/0302/c14657-32363748
杨毅周：维护台胞权益 助力台胞更好融入国家发展大局/n1/2022/0309/c14657-32370886.html
台盟中央副主席：解决好台青登陆创业就业“水土不服”问题/n1/2022/0308/c14657-32369278.html
江尔雄：我很喜欢《我们同唱一首歌》，欢迎台青来大陆追梦筑梦圆梦/n1/2022/0308/c14657-32369272
两岸同胞富春江畔踏青 携手共绘现代富春山居图/n1/2022/0307/c14657-32368128.html
台湾团体：发扬“二·二八”爱国爱乡精神，携手实现民族复兴/n1/2022/0226/c14657-32360328.html
国台办发言人介绍苏州新冠肺炎疫情涉及台企台胞有关情况/n1/2022/0223/c14657-32358033.html
国台办：台湾同胞以各种方式参与北京冬奥会是“两岸一家亲”的真实写照/n1/2022/0223/c14657-3235
大陆市场助力台湾工具机业逆势成长/n1/2022/0222/c14657-32357016.html
江尔雄：我很喜欢《我们同唱一首歌》，欢迎台青来大陆追梦筑梦圆梦/n1/2022/0308/c14657-32369272
揭开“爱台”与“挺台”的谎言—评民进党“舔美”新闹剧/n1/2022/0307/c14657-32368150.html



: /wm/116 “/wyz/81 ” /wyy/81 的/ude1/58 台湾/ns/43 北京/ns/32 :/w/31 http/x/31 ，/wd/27
中国/n/26 、/wn/24 增/v/20 在/p/20 新冠肺炎/n/18 （/wkz/17 ）/wky/17 确诊/v/17 例/q/17
新/d/17 委员/n/17 冰雪/n/16 冬残奥会/n/16 大陆/n/16 一/m/16 国际/n/15 新/a/15 冬奥会/n/14
病例/n/14 会/v/14 为/p/14 对/p/14 安全/n/13 国家/n/13 国/n/13 人/n/13 高/a/12 好/a/12
人民/n/12 -/w/12 代表/n/11 更/d/11 出/vf/10 台/q/10 《/wkz/10 国务院/nt/10 安/ag/10 委/ng/1
》/wky/10 观察/v/10 两岸/n/10 我/n/10 碳/n/10 助力/b/9 台胞/n/9 冬奥/n/9 梦/n/9 发展/v/9
有/vyou/9 公平/a/9 让/v/9 大/a/9 科技/n/9 春/tg/9 共/d/9 三/m/8 问题/n/8 青/a/8 -/w/8 和/cc/
司法/n/8 是/vshi/8 武警/n/8 图/n/8 台/n/8 中/b/8 战/vg/7 精神/n/7 特/ag/7 等/udeng/7 奥/b/7
多/a/7 从/p/7 之/uzhi/7 日本/nsf/7 个/q/7 ！/wt/7 台办/n/7 市场/n/7 残/vg/7 某/rz/6 黄河/ns/6
加快/v/6 年/qt/6 障碍/n/6 防凌/vi/6 元宵/n/6 反/vi/6 完善/v/6 服务/vn/6 显/v/6 无/v/6 7/m/6
到/v/6 正义/n/6 五/m/6 全球/n/6 向/p/6 世界/n/6 推动/v/6 前/f/6 上/f/6 体育/n/6 已/d/6
刘梦涛/nr/6 2月/t/6 起/q/6 彰/ag/6 后/f/5 解决/v/5 同/p/5 部队/n/5 锻造/v/5 守护/v/5 网/n/5
我们/rr/5 了/ule/5 减/v/5 线/n/5 发布/v/5 双/m/5 圆/vg/5 提供/v/5 群众/n/5 做/v/5 党/n/5
将/d/5 绿色/n/5 近期/t/5 实战/n/5 登陆/vi/5 组/n/5 质量/n/5 同胞/n/5 以/p/5 方式/n/5 应/v/5


病例
军事
大陆
凸显
国家
完善
尖兵
发布
冬奥盛会
女性
打造
官兵
科技支撑
中央
举行
特战队员
我国
计划

代码和ppt放在了github上



https://github.com/Andrew0613/Big-Data-Analysis-and-Application

Thank you !