# 预训练卷不动，可以卷输入预处理啊！

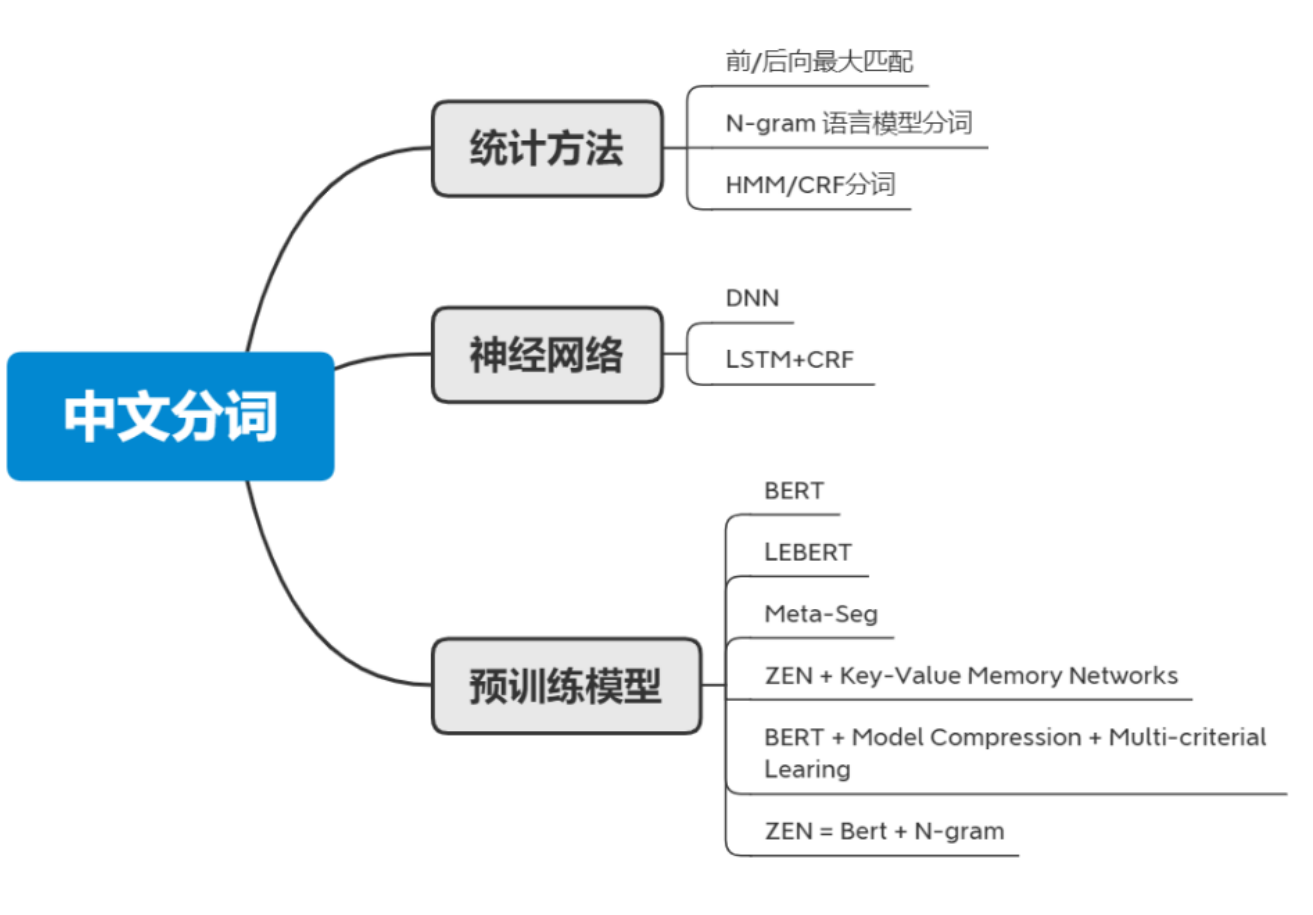
原创 德志 夕小瑶的卖萌屋 2021-07-13 12:37

文 | 德志
编 | 小戏

目前伴随着预训练预言模型的兴起，越来越多的 NLP 任务开始脱离对分词的依赖。**通过 Fine-Tune Bert 这类预训练预言模型，能直接在下游任务上取得一个很好的结果**。同时也有文章探讨中文分词在神经网络时代的必要性。对于分词任务本身也是如此。

那中文分词这个任务有还意义吗？或者换句话说中文分词是不是一个已经解决的任务。**那么接下来笔者将会带大家梳理目前分词的研究方向和进展。**

本文的思维导图如下图所示。其中，“统计方法”和“神经网络”两部分会简单介绍一下早期的传统做法，熟悉的同学可以直接跳过。**主体在最后的“预训练模型”部分，会带大家梳理一下2020年以来的最前沿的一些中文分词工作**。



▲本文思维导图

## 任务描述

分词任务相信大家都不陌生了，**其实就是给定一个句子，让后将一个句子切分成一个个的基本词**。

例如：'上海浦东开发与建设同步' → ['上海', '浦东', '开发', ‘与', ’建设', '同步']。

对这个任务的解法也有很多中，比如最开始的**前/后向最大匹配**，后来的也有 **N-gram语言模型**，**HMM/CRF** 的分词方法，再到现在的**基于深度学习的端到端的分词方法**。总而言之，分词的方法也是跟着时代是在不断进步的。

## 前浪们：统计方法

对于分词这项任务最早的方案是**依靠词典匹配**的方式，到后来**利用统计信息进行分词**，最后采用了**序列标注**的方案进行分词。这些方案的代表方法有：

1. **前/后向最大匹配**：其朴素思想就是利用词表采用贪心的方式切分出当前位置上长度最大的词作为分词结果返回。
2. **N-gram 语言模型分词**：其思想在于利用统计信息找出一条概率最大的路径。一般需要很大量的数据才能统计的很准。
3. **HMM/CRF分词**：把分词当做一个序列标注问题。序列单元是字，序列标签有B,M,E,S，分别代表词首，词中，词尾和单词。

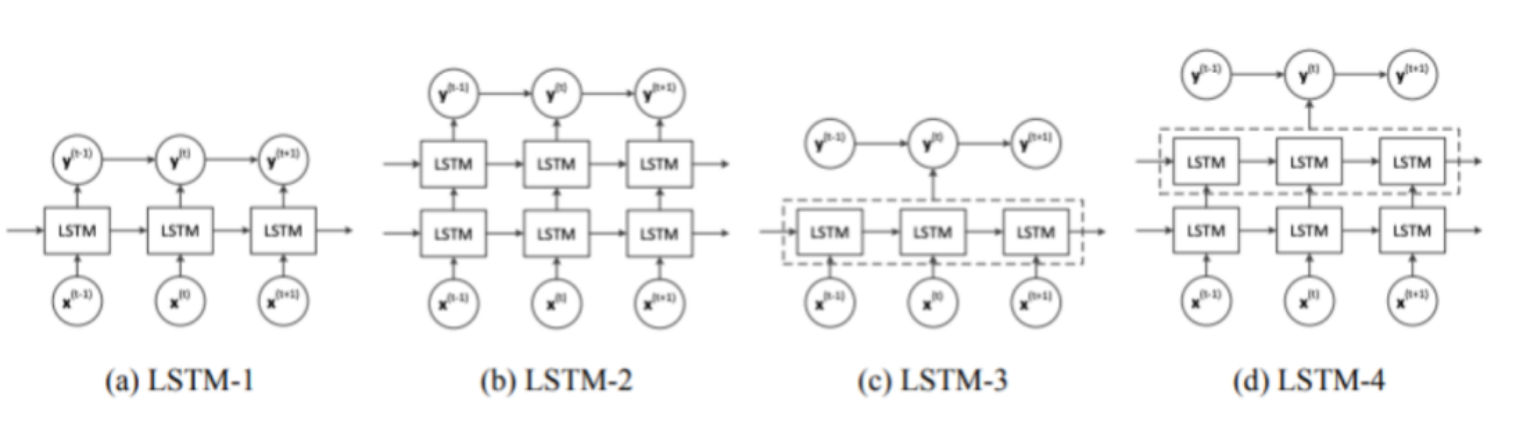
前浪们的方法就不赘述太多了，这些方式都或多或少存在一定的局限性，当然，这些方法显著的优势是它们速度都很快。

## 中浪们：神经网络

步入到深度学习时代，开始涌现出形形色色**利用神经网络的分词方式**。一个朴素的方案是，**给定一个中文的句子，$X=\{x_1,x_2..,x_t\}$ ,输出的一个 Label 序列 $Y=\{y_1,y_2..,y_t\}$**。

Label 序列是由{B，M，E，S}组成。其中，B 为词的开头，M 代表词的中间，E 为词的结尾，S 指的是单字。**这种方案首先将句子切分成单字输入到模型中，通过序列标注的形式进行学习。**

之后，中浪们开始**采用了各种模型去提取字符特征，然后利用 CRF 进行序列标注的学习**。比较典型的方案是 **LSTM+CRF** 的方式。

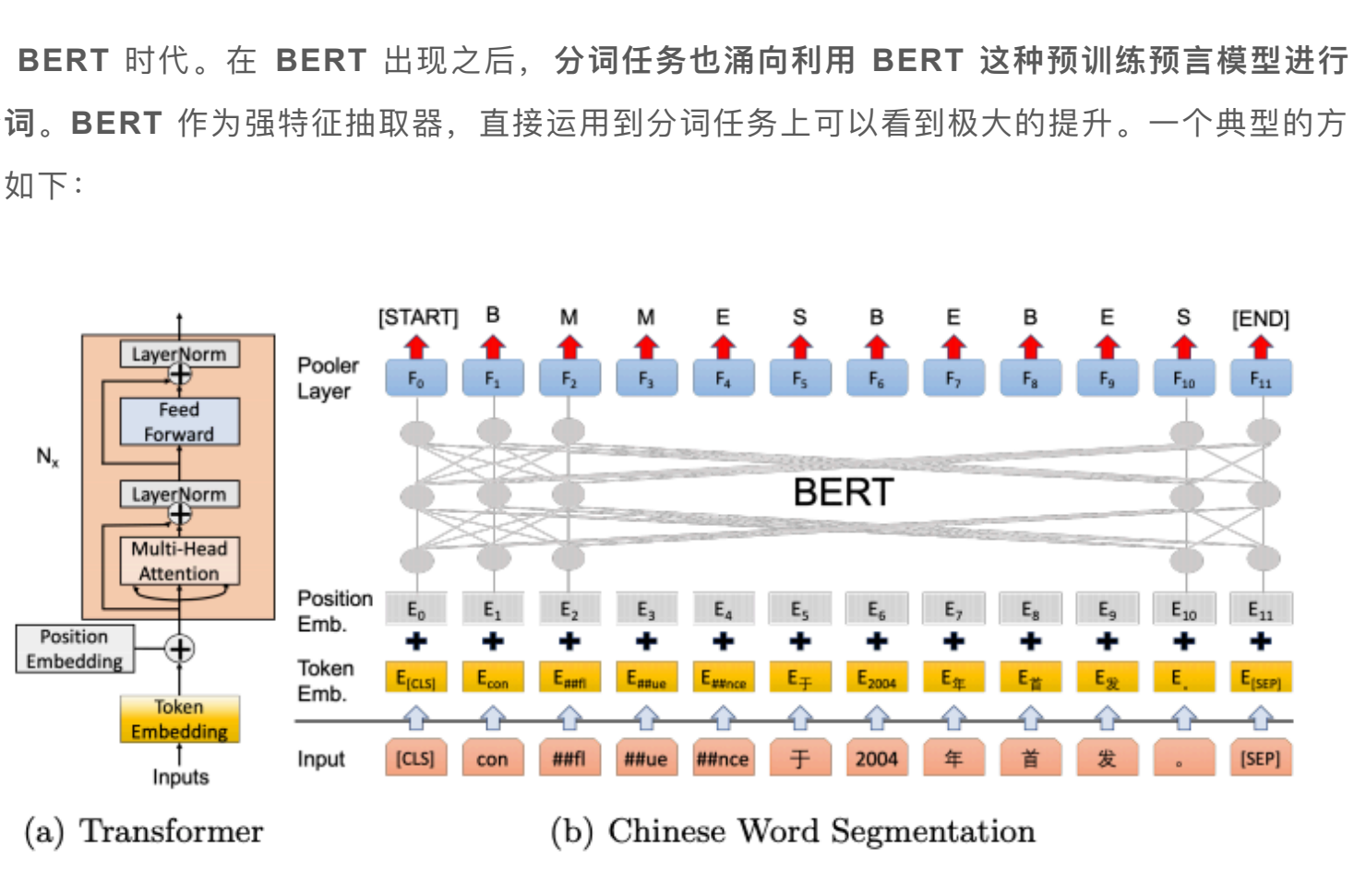


(a) LSTM-1 (b) LSTM-2 (c) LSTM-3 (d) LSTM-4

**LSTM** 的优势在于能够**保留之前的有效信息**，以及**减少窗口的限制**。对比传统方法而言，基于 NN 的方法效果好且对于歧义词和未登陆词有优势，虽然在速度上不如传统模型。

## 后浪们：预训练时代

后 **BERT** 时代。在 **BERT** 出现之后，**分词任务也涌向利用 BERT 这种预训练预言模型进行分词。BERT** 作为强特征抽取器，直接运用到分词任务上可以看到极大的提升。一个典型的方式如下：



(a) Transformer (b) Chinese Word Segmentation

**Fig. 2.** Architecture of Transformer and BERT for the CWS task.

此时可能大家就会有一个疑问：**是不是对于分词任务，使用一个 BERT 就好了，还有什么可以研究的方向吗**？而我个人总结有如下几个研究方向。

1. **如何通过不同粒度标准的分词预料联合预训练，让分词能够通过某些简单的控制能够适应不用的分词场景？**
2. **如何在 NN 模型中融合自定义词典的功能**？比如输入时融入额外的 **Embedding**；
3. **如何将 NN 分词框架和外部知识结合**？比如如何结合外部的字典树等问题；
4. **如何大的 NN 模型蒸馏成一个小的模型**？即如何将大的 **BERT** → 小的 **CNN/LSTM/BERT**？由于分词模型的场景对性能要求很高，因此把深度模型的速度提升是目前急需解决的问题。

对于以上的几个热门方向分别有如下代表方案：

### LEBERT（2021 ACL）

**LEBERT** 的主要的方案是**在输入的时候需要采集句子中的字符-词语 pair ，通过词典匹配（字典树）**——这个词典是由预训练的 **Word-Embeding** 的词组成的——**然后通过 Lexicon Adapter 往 BERT 中注入词特征。采用字向量+加权求和得到融合后的词向量。词向量本身是通过额外训练的。**

通过下图可以明显的看到整个 **LEBERT** 的整体结构。给定一个句子[美国人民],对于每个句子中的字都会有一个字符-短语的 pair，"美"->[美国，美国人,<pad>],<pad>是为了对齐。

然后在求和的时候作者设计了 **Lexicon Adapter** 对字向量和短语 pair 的词向量进行求和，剩下的就和原生 **BERT** 一致了。额外的 **Word-Embeding** 则是采用了**腾讯 AI-Lab** 开源的词向量。

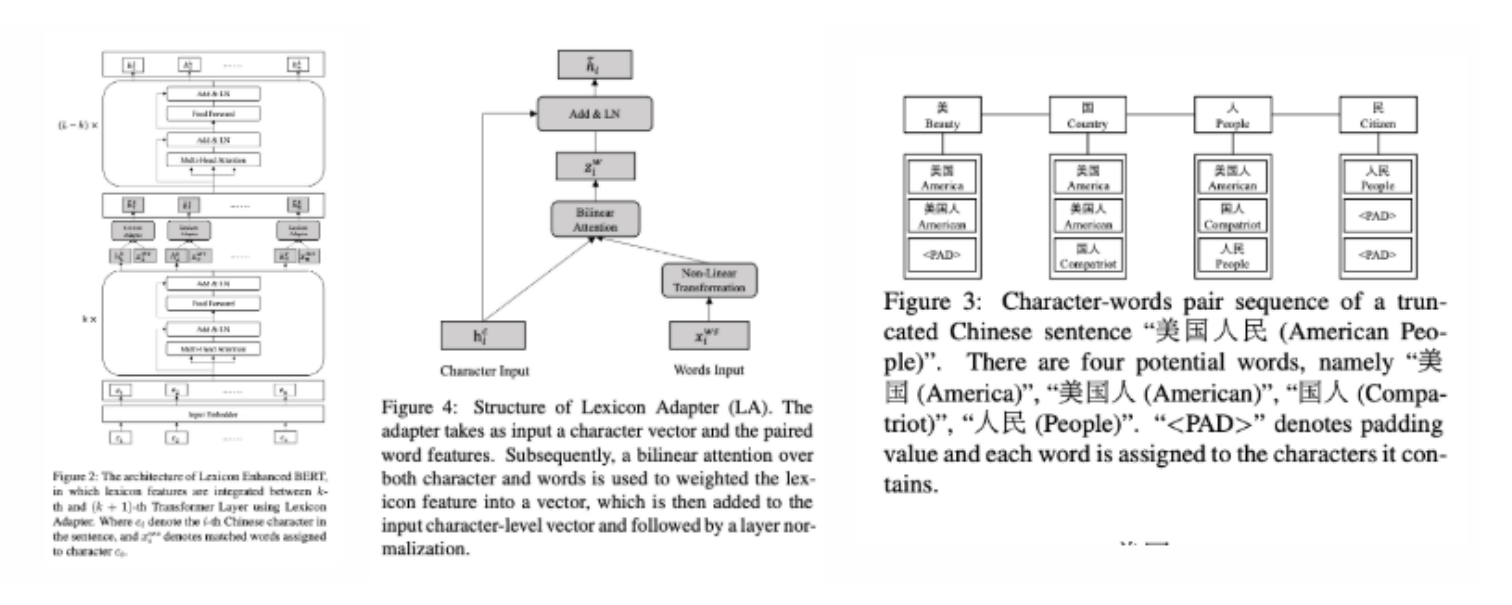


Figure 2: The architecture of Lexicon Enhanced BERT, in which lexicon features are integrated between k-th and (k + 1)-th Transformer Layer using Lexicon Adapter. Where $c_i$ denote the i-th Chinese character in the sentence, and $x_i^{ws}$ denotes matched words assigned to character $c_i$.

Figure 4: Structure of Lexicon Adapter (LA). The adapter takes as input a character vector and the paired word features. Subsequently, a bilinear attention over both character and words is used to weighted the lexicon feature into a vector, which is then added to the input character-level vector and followed by a layer normalization.

Figure 3: Character-words pair sequence of a truncated Chinese sentence “美国人民 (American People)”. There are four potential words, namely “美国 (America)”, “美国人 (American)”, “国人 (Compatriot)”, “人民 (People)”. “<PAD>” denotes padding value and each word is assigned to the characters it contains.

> 论文题目：
> **Lexicon Enhanced Chinese Sequence Labelling Using BERT Adapter**
> 论文链接：
> *https://arxiv.org/pdf/2105.07148.pdf*

### Meta-Seg(2021 NAACL)

**Meta-Seg** 构建了**第一个多粒度的分词预训练语言模型。并通过元学习的方式进行多粒度的预训练**。其衍生的姊妹篇文章则是通过**引入 Bigram+额外的损失函数来构建多粒度的分词**。共同的做法是输入端增加来引入是哪种分词粒度的信息，而不同的是，**Meta-Seg目标是通过元学习让模型学到不同数据集下的分割标准。**

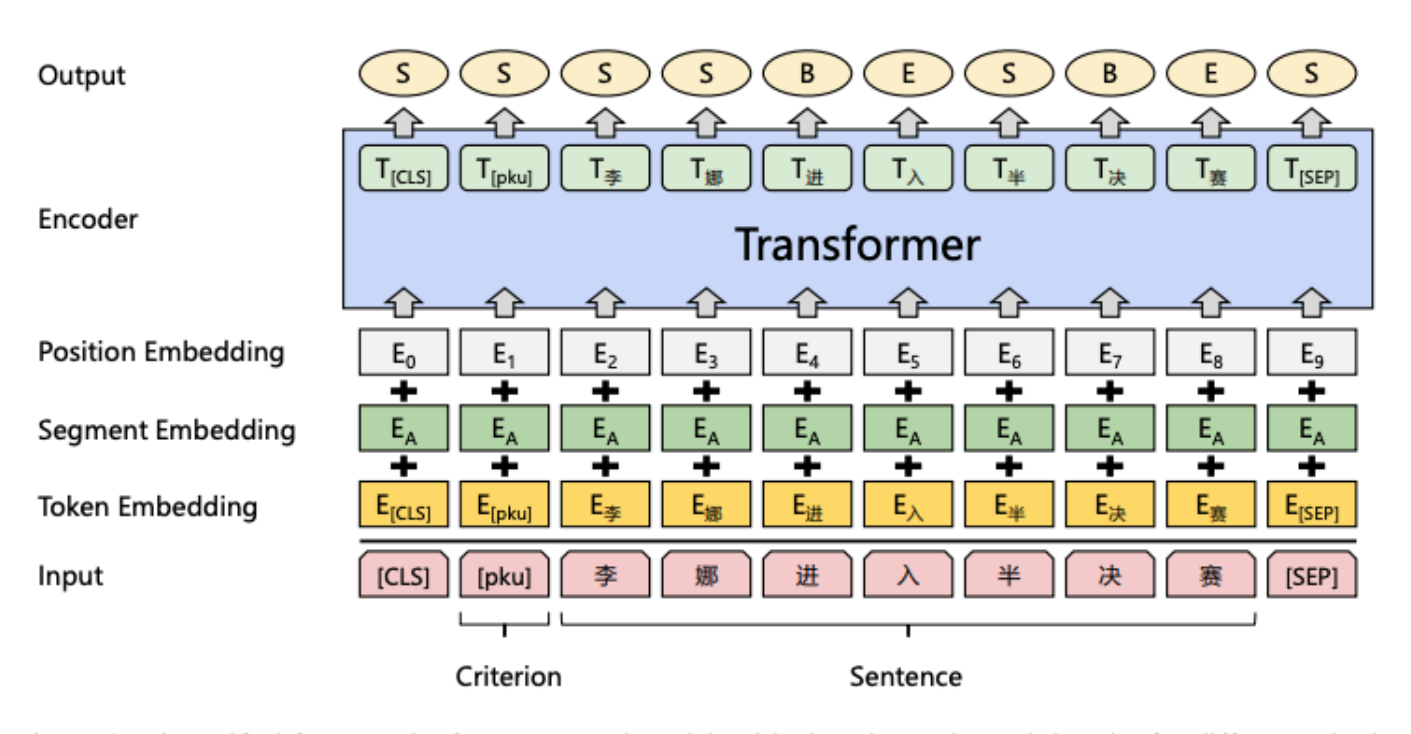


Figure 1: The unified framework of our proposed model, with shared encoder and decoder for different criteria. The input is composed of criterion and sentence, where the criterion can vary with the same sentence. The output is a corresponding sequence of segmentation labels of given criterion.

> 论文题目：
> **Pre-training with Meta Learning for Chinese Word Segmentation**
> 论文链接：
> *https://arxiv.org/pdf/2010.12272.pdf*

### ZEN + Key-Value Memory Networks(2020 ACL)

**ZEN + Key-Value Memory Networks** 一文的核心思想是在传统的 **CWS** 模型上加入 **Memory Networks** 缓解OOV的问题。

**Encoder** 可以是任意的网络（BERT/ZEN），**Decoder**部分则是 **Softmax** 或者 **CRF** 。核心是 **Wordhood Memory Networks**。

**Wordhood Memory Networks 可以认为是一种 Key-Value 的存储结构**。该方法的核心在**于首先构建一个 N-gram 的词表。然后对于每一汉字而言，所有得到所有包含该字的 N-gram 作为Key，Value则是同样的一个列表，表示的是字在 N-gram 中的位置。**

给定一个词“民”，Memory 的 Key 为[民，居民。民生，居民生活]，Value则是[S,E,B,I]，分别代表了“民”在 **N-gram** 中的位置。**然后用民的 Embeding 对 N-gram 的词进行点乘取 Softmax 就得到相关性。**

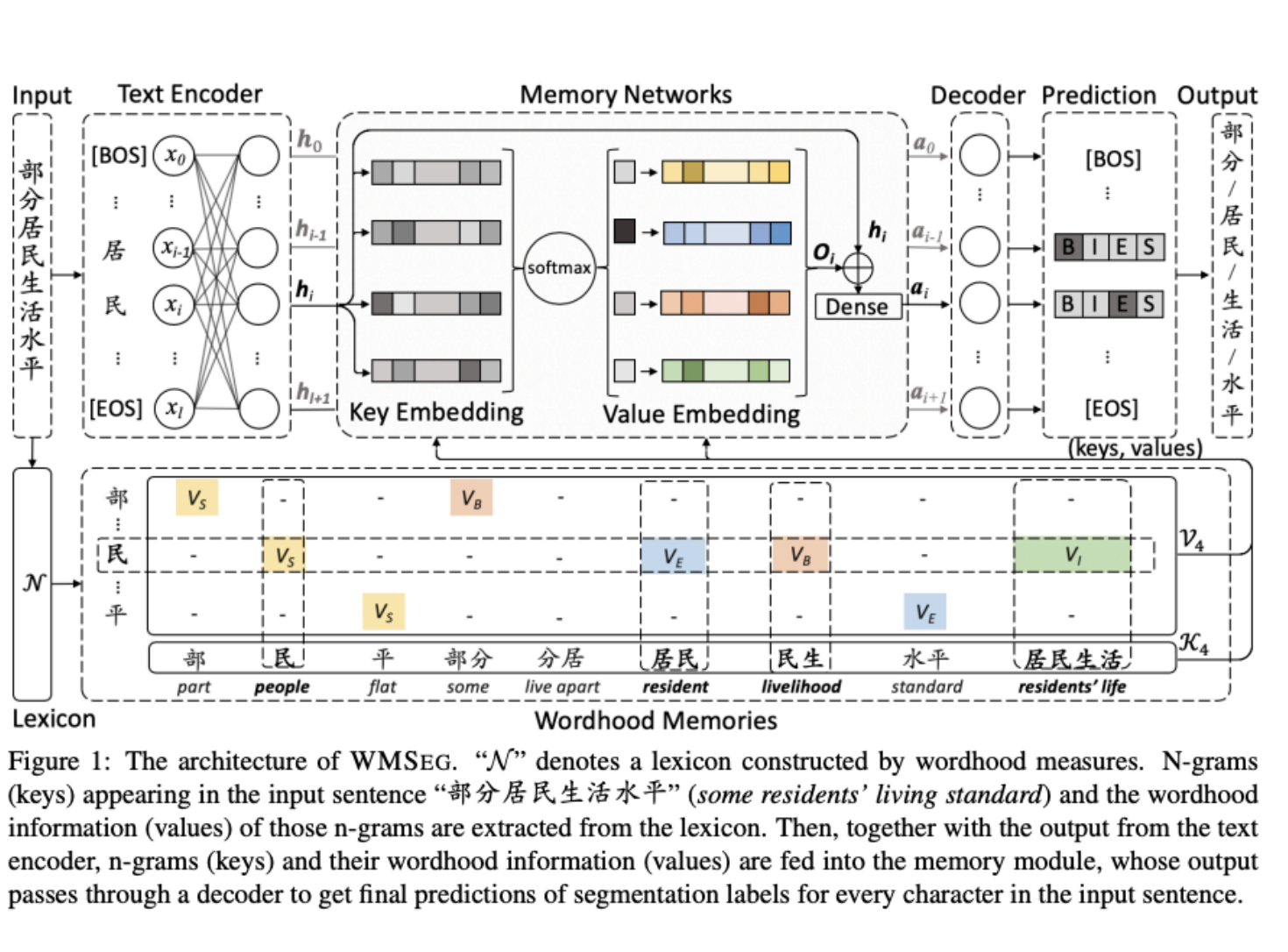


Figure 1: The architecture of WMSEG. “$\mathcal{N}$” denotes a lexicon constructed by wordhood measures. N-grams (keys) appearing in the input sentence “部分居民生活水平” (*some residents’ living standard*) and the wordhood information (values) of those n-grams are extracted from the lexicon. Then, together 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.

> 论文题目：
> **Improving Chinese Word Segmentation with Wordhood Memory Networks**
> 论文链接：
> *https://aclanthology.org/2020.acl-main.734v2.pdf*

### BERT + Model Compression + Multi-criterial Learing(2020 COLING)

**BERT + Model Compression + Multi-criterial Learing 的想法非常简单粗暴，由于分词标注的主观性导致了现有数据集在分词粒度上会有分歧，所以想利用某种方式捕获粒度不同且能够利用共同基础知识。**

方案很简单，**构建一个共有的影层学习共有知识，构建一个私有隐层破获独特性，然后将两个层的结果加起来进行标签预测**。而模型压缩这块还是使用了蒸馏的方式，蒸馏了一个 3 层的小 **BERT** 。**Student** 的学习是通过 **Teacher-Students** 损失+标签损失学习。

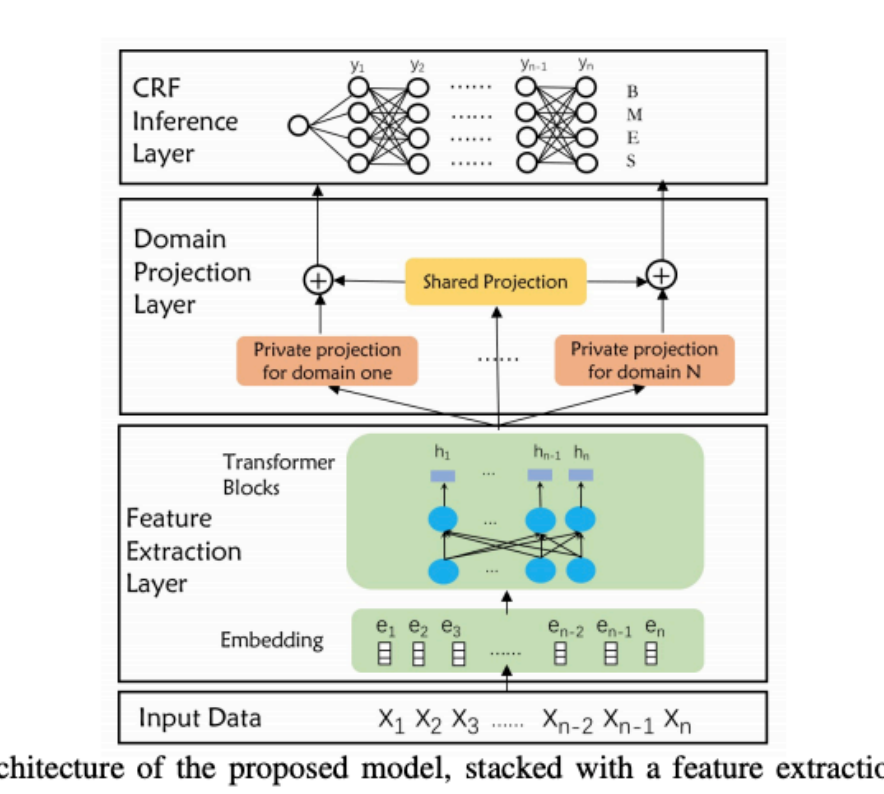


Figure 1: The architecture of the proposed model, stacked with a feature extraction layer, a domain projection layer and a CRF tag inference layer.

> 论文题目：
> **Unified Multi-Criteria Chinese Word Segmentation with BERT**
> 论文链接：
> *https://arxiv.org/pdf/2004.05808.pdf*

### ZEN = Bert + N-gram(2020 EMNLP)

**ZEN = Bert + N-gram 引入 N-gram 编码方式，方便模型识别出可能的字的组合。N-gram** 的提取分成两步，首先通过语料生成 **N-gram** 词表，然后通过此表生成 **N-gram Matrix**。

**N-gram Embedding 的方式则是和 BERT 的 Embeeding 一致**。字 **Embedding** 和 **N-gram Embedding** 的结合方式则是直接做了矩阵相加。

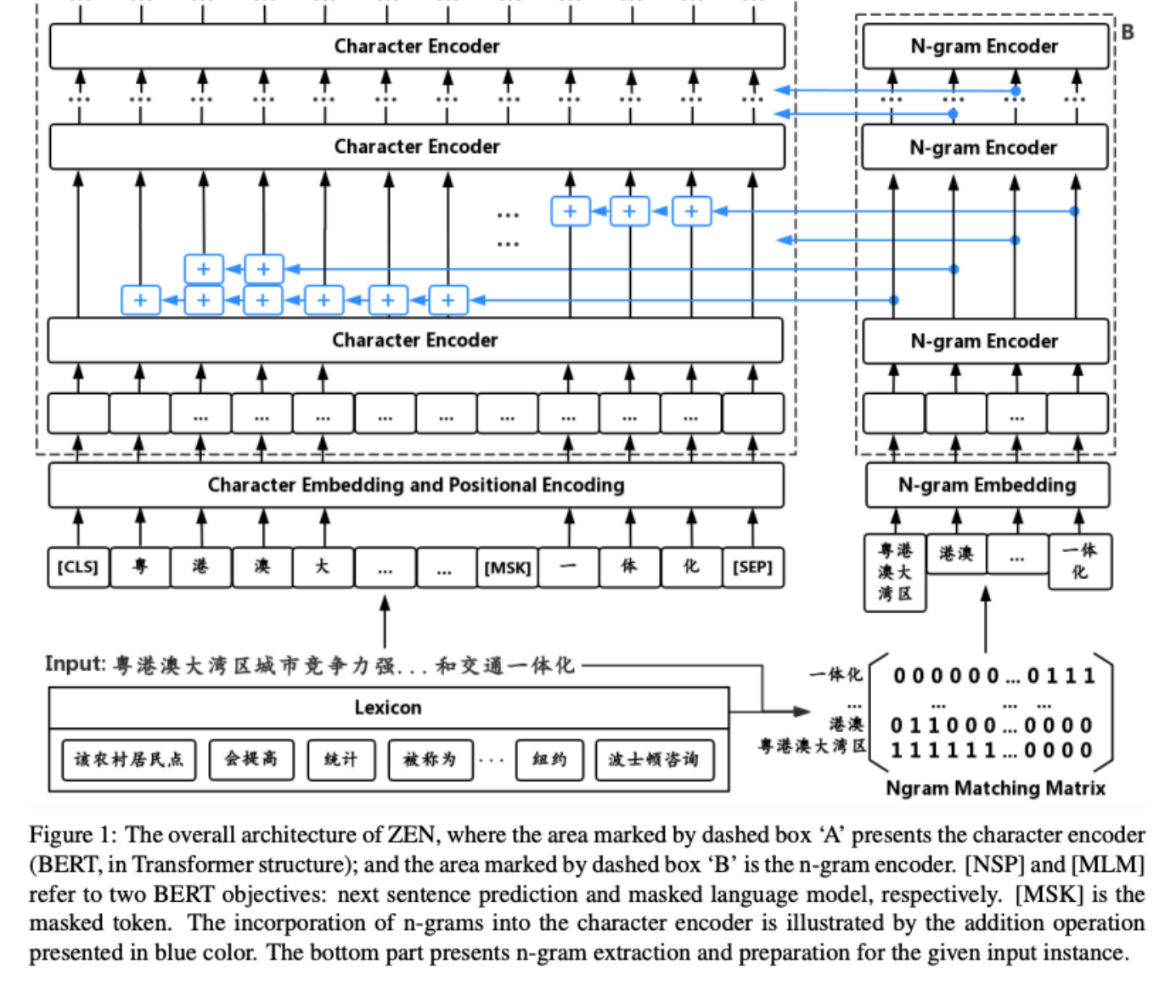


Figure 1: The overall architecture of ZEN, where the area marked by dashed box ‘A’ presents the character encoder (BERT, in Transformer structure); and the area marked by dashed box ‘B’ is the n-gram encoder. [NSP] and [MLM] refer to two BERT objectives: next sentence prediction and masked language model, respectively. [MSK] is the masked token. The incorporation of n-grams into the character encoder is illustrated by the addition operation presented in blue color. The bottom part presents n-gram extraction and preparation for the given input instance.

> 论文题目：
> **ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations**
> 论文链接：
> *https://arxiv.org/pdf/1911.00720.pdf*

## 目前 SOTA 排行

### 常用的数据集

| Crops | Test Size(Tokens/Types) | Train Size |
|---|---|---|
| Traditional Chinese | - | - |
| AS | 122K / 19K | 5.45M |
| CityU | 104K / 13K | 1.46M |
| Simplified Chinese | - | - |
| PKU | 41K / 9K | 1.1M |
| MSR | 107K / 13K | 2.37M |

### 采用 F1 的评价标准

| Model | AS | CITYU | MSR | PKU |
|---|---|---|---|---|
| LEBERT | - | - | 98.6 | 96.9 |
| Lattice-BERT-base | - | - | 98.6 | - |
| METASEG (bert + meta-learning + multi-criterial) | 97.0 | 98.2 | 98.5 | 96.9 |
| bert-12-layer + multi-criterial learing–teacher | 97.0 | 97.8 | 98.5 | 97.3 |
| ZEN + key-value memory networks | 96.6 | 97.9 | 98.4 | 96.5 |
| ZEN | - | - | 98.3 | 96.3 |
| BERT + model compression + multi-criterial learing–student | 96.6 | 97.6 | 97.9 | 96.6 |
| BERT-base | - | 97.9 | 96.2 | |
| BiLSTM-CRF | - | - | 96.4 | 95.7 |

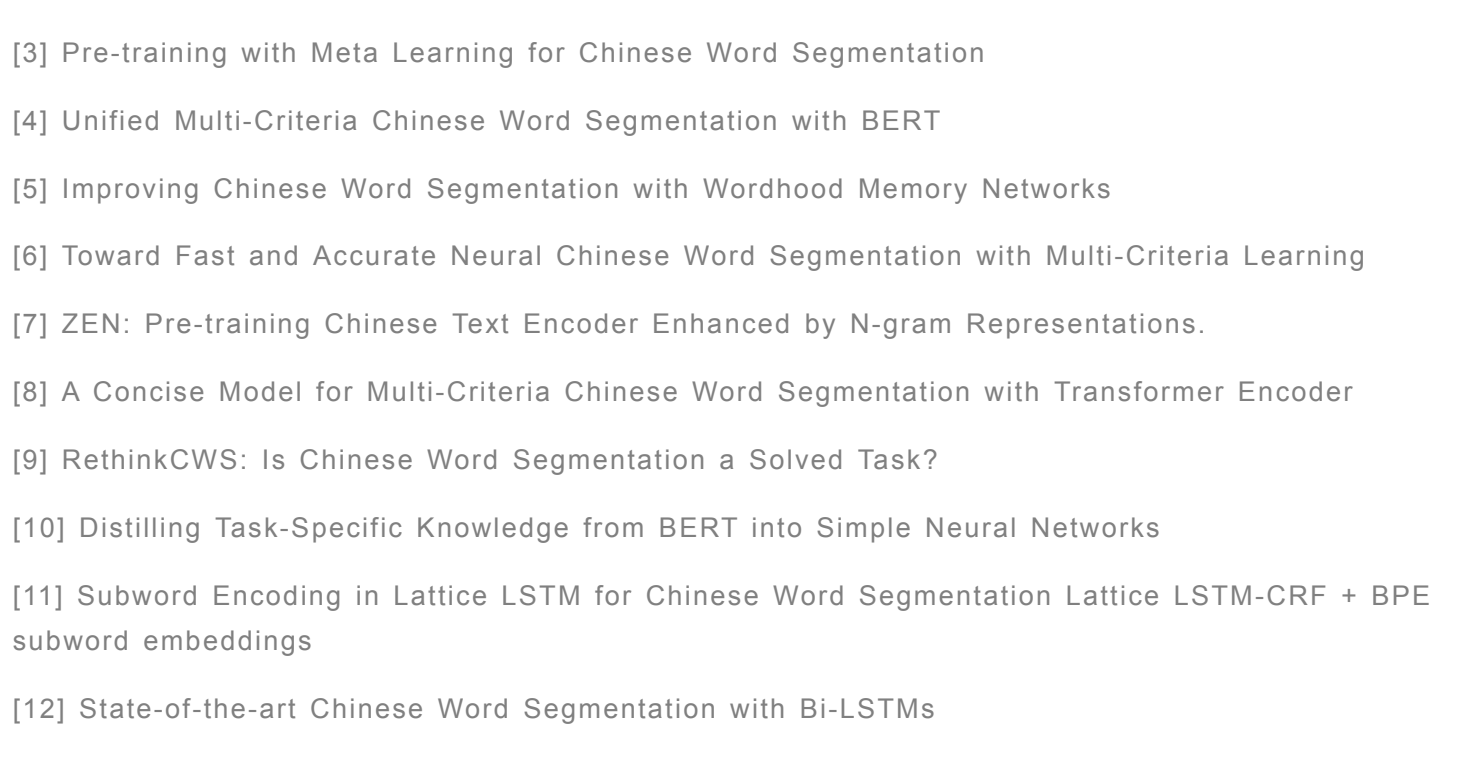
## 总结

本文回顾了分词的发展历程，以及目前的研究热点方向。**总的来说分词任务其实发展至今可以看到在公开数据集上已经有了很好的效果，但是在实际运用上切词的效果总是没那么让人满意**。其主要问题有：

1. **实际使用上用户比较关注效率问题，比如如何提升 NN 模型的效率？**
2. **每天有大量的新词产生，对于 OOV 的问题如何更有效的解决？**
3. **词的界限不明确，大家对分词的标准不统一。**

这三点导致了目前实际使用中分词效果大大折扣。未来分词还有很多方向需要大家探索，在 **RethinkCWS** 一文中也有很多对中文分词的目前看法，感兴趣的大家可以去参考查阅一下。

> 论文题目：
> **RethinkCWS: Is Chinese Word Segmentation a Solved Task?**
> 论文链接：
> *https://arxiv.org/pdf/2011.06858.pdf*



## 参考文献

[1] Lexicon Enhanced Chinese Sequence Labelling Using BERT Adapter

[2] Lattice-BERT:Leveraging Multi-Granularity Representations in Chinese Pre-trained Language Model

[3] Pre-training with Meta Learning for Chinese Word Segmentation

[4] Unified Multi-Criteria Chinese Word Segmentation with BERT

[5] Improving Chinese Word Segmentation with Wordhood Memory Networks

[6] Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning

[7] ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations.

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

[9] RethinkCWS: Is Chinese Word Segmentation a Solved Task?

[10] Distilling Task-Specific Knowledge from BERT into Simple Neural Networks

[11] Subword Encoding in Lattice LSTM for Chinese Word Segmentation Lattice LSTM-CRF + BPE subword embeddings

[12] State-of-the-art Chinese Word Segmentation with Bi-LSTMs

[13] Neural Networks Incorporating Dictionaries for Chinese Word Segmentation.

[14] Adversarial Multi-Criteria Learning for Chinese Word Segmentation

[15] Long Short-Term Memory Neural Networks for Chinese Word Segmentation BiLSTM-CRF

[16] Ambiguity Resolution in Chinese Word Segmentation

[17] 中文分词十年回顾

[18] 中文分词十年又回顾