北京理工大學

# 计算机新技术专题论文

## 中文分词发展与技术研究

**A survey of Chinese word segmentation development and techniques**


学　　院：计算机学院

专　　业：计算机科学与技术

学生姓名：赖昱行

学　　号：1120192236

指导教师：黄河燕


2021 年 11 月 19 日

# 中文分词发展与技术研究

## 摘　要

在中文自然语言处理任务中，中文分词属于最基础的任务，长期以来是其他任务的前提和基石。回顾中文分词技术的发展并进行归纳和总结，可以为后来的研究提供参考。本文整理了中文分词的必要性和关键问题，梳理了机械分词、机器学习和深度学习的典型算法和模型，并给出了优劣势的分析。同时选取了 31 篇来自各类期刊会议的中英文文献进行观点佐证。结果显示，目前以深度学习为基础的模型最为突出，在已有的成果上进一步提高了中文分词的准确率。本文未能对近年的优秀模型进行深入的讨论和比较。中文分词领域在未来仍有集成算法、联合模型、迁移学习、多模态等领域值得研究和开拓。

**关键词：中文分词；机器学习；深度学习；语义理解**

# A survey of Chinese word segmentation development and techniques


## Abstract

In Chinese natural language processing tasks, Chinese word segmentation is the most basic task, and has long been the premise and cornerstone of other tasks. Reviewing the development of Chinese word segmentation technology and making a summary and conclusion can provide guidance for future research. This article sorts out the necessity and key issues of Chinese word segmentation, gives a brief introduction to the typical algorithms and models of mechanical word segmentation, machine learning and deep learning, and gives an analysis of the advantages and disadvantages. At the same time, 31 Chinese and English papers from various journals and conferences were referred to support my opinions. The results show that the current deep learning-based model is the most prominent, and it has further improved the accuracy of Chinese word segmentation compared with the existing results. This article hasn't yet conducted in-depth discussion and comparison of the excellent models in recent years. In the field of Chinese word segmentation, there are still areas worthy of research and development in the future, such as integrated algorithms, joint models, transfer learning, and multi-modality.


**Key Words: Chinese Word Segmentation; Machine Learing; Deep Learning; Semantic Understanding**

# 目 录

# 第 1 章 中文分词的必要性和问题

## 1.1 必要性

在自然语言处理中，按照对象处理的颗粒度，可分为文本识别，词法分析，句法分析，语义分析和篇章分析。文本识别往往是对语音或图像进行处理，通过语音识别或者光学字符识别的方式提取出文本。事实上，在现实生活中，我们有大量的语料是直接以文本的形式呈现的，无需进行额外的识别。故而词法分析可以说是最基本的底层任务，是后续高级任务的基石。在流水线系统中，如果词法分析出错，往往会很大程度上波及后续任务；而在近几年的端到端联合模型中，词法分析的任务也会提升模型在其他任务上的效果[1]。由此可见词法分析的重要性和必要性。

词法分析的主要任务又分为中文分词，词性标注和命名实体识别。在传统的，同时也是主要的流水线系统中，中文分词又作为后两个任务的基础。这是由中文的孤立语性质决定的。语序是汉语句法结构中的主要表达手段，而分词的正确性决定了语序，也就是结构层面的合理性[2]。对于大部分印欧语系的语言来说，它们的文本具有天然的分词性质，而中文文本没有明显的边界，所以我们需要在词与词之间做出明确的分割标记，也就是对中文文本这一连续字符串进行序列切分。

## 1.2 问题

学术界和工业界对中文分词的探讨已经有了比较可观的结果，在SIGHAN2005[3]等权威数据集上进行的测试已经可以达到95%的F值[4]。然而，剩下的5%却很难进行突破，主要原因在于自然语言高度灵活，对人来说，可能无法切身体会到语言的复杂性，但对机器来说，充分理解语言是一件非常困难的事情。事实上，目前认为中文分词有以下关键难点。

### 1.2.1 缺乏统一的分词标准

在进行分词时，由于事实标准，自定义标准或者特定领域问题标准的不同，容易导致分词结果不统一。例如经典的1998年人民日报语料库[5]中，作者便提到辅助专家针对新闻语料的特点进行了标注。不同的文本背景下，不同的团队有着不同的视角和研究问题，往往也难以制定统一的标准。同时，哪怕有着统一的分词标准，也难免错误的人工标注出现，刘江[6]曾针对这一问题进行了分词一致性检验的研究。而人工分

词的语料库中的错误，最后会反映在中文分词的效果上，特别是基于统计学原理的分词方法。

### 1.2.2 切分歧义

自然语言含有大量歧义，中文也不例外。在不同的语境下，同一个词可能表现为多种含义。比如多义词，必须通过上下文乃至说话者的身份才能判断。又或者是利用语义双关制造幽默、讽刺效果的词语。在中文文本中，具有歧义的词显得尤为多。仅以1998年人民日报语料中的兼类词为例，本文按词频进行排序，统计了前$total$个词中具有两个及以上词性的词语数量及占比。

表 1-1 兼类词统计表

| $total$ | 兼类词数 | 占比 |
|---|---|---|
| 100 | 59 | 0.59 |
| 200 | 115 | 0.575 |
| 500 | 265 | 0.53 |
| 1000 | 509 | 0.509 |
| 5000 | 1969 | 0.3938 |
| 10000 | 3207 | 0.3207 |

### 1.2.3 未登录词

中文同任何自然语言一样都是不断发展的，其中比较显著的便是词语的变化。虽然存在普通话、简体字等规范，但是仍然有大量的新词出现，或者给旧词汇赋予新的含义。就如同古代汉语和现代汉语的差别巨大，而中文也在不断吸收英语、日语中的外来词汇。在互联网时代下，交流的便捷性更是加速了这一过程，每年不断有流行词诞生，又有流行词消失；在各个领域，新的专业名词、专有名词也在不断涌现[7]。这样的特点给词语的识别，进而给中文分词带来了不小的困难。为了处理这一难题，在第六届MUC会议上正式提出了命名实体[8]的概念，将命名实体识别这个任务单独划分出来进行研究。

### 1.2.4 简略性

在正式的文本中，语言往往详实具体，便于利用规则进行分析，但是在对话等非

正式的场合，语言往往偏向于简洁、干练，省略掉了大量信息。这些信息不仅是上下文中可能蕴含的内容，更可能是基于人的身份、对话场景、文化背景等因素。对于这样的非正式文本，人工切分尚且困难，对于机器而言便更不是一件容易的事情。

# 第 2 章 传统分词方法

传统的中文分词模型算法根据其原理和特点可以大致分为两类——基于词典的匹配方法和基于统计的分词方法。

## 2.1 基于词典的匹配方法

这种方法又称机械分词，是一种利用规则进行分词的方法。这种算法使用一定的策略，将文本作为字符串，对一个已知的、充分大的词典中的词进行匹配。故而策略是否合理，词典是否足够丰富便显得非常重要。一般而言，此类方法算法简单，分词效率高，但切分歧义问题较为严重。常见的机械分词方法有以下几种。

### 2.1.1 正向最大匹配法（FMM）

其基本思想是，对于待切分的一段文本，以贪心的方式选取当前位置开始的数个字符与词典进行匹配，并选取其中最长的可匹配词作为匹配结果，并以此为依据进行切分。其分词原理可解释为，单词的颗粒度越大，所能表达的含义越确切。

### 2.1.2 逆向最大匹配法（RMM）

该方法思想与FMM基本类似，不过在进行切分时，从文本末尾开始，并且取字符的方式变为从右到左，同样以贪心的方式匹配最长的词。有文章[9]指出，在统计的方法下，该方法的分词错误率明显低于FMM。

### 2.1.3 双向最大匹配法

该方法同时采用FMM和RMM，并取二者分词结果中词数较少者。虽然词数较少并不意味着切分正确，在实际应用中错误率更低。

### 2.1.4 最短路径分词法

最短路径分词法首先将文本中所有的字符都看作单独的个体，并在相邻的字符间连有向边。而后，从每一个字符开始向后拼接，用拼接后的字符串对词典中的词进行匹配，若能匹配到，则向末尾字符添加一条有向边。词图构建完成后，可以发现这

样的图为DAG。由此，我们可以采用动态规划或者贪心算法对从首字符到尾字符的最短路径进行求解，并取此最短路径为最终分词结果。

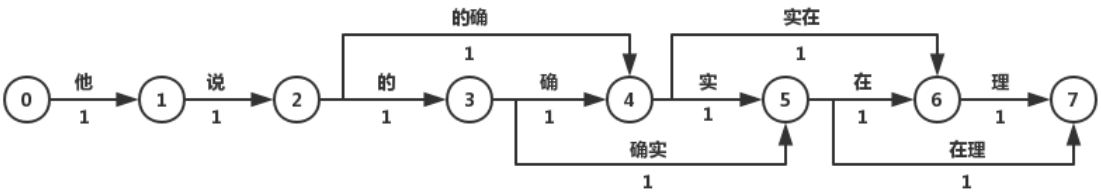


图 2-1 最短路径分词法

最短路径分词法相比前述三种最大匹配法在理论复杂度上一致，在分词效果上却有较大的提升。基于这种方法，还提出了最大概率分词法和N-最短路径分词法作为改进，区别在于有向边的边长和保留路径的数量。有文章研究结果显示，基于N-最短路径的统计粗分模型[10]在效率上和分词准确率、召回率上均取得了突出表现。

## 2.2 基于统计的分词方法

基于统计的分词方法一般基于机器学习，需要进行特征的人工设计。不同于机械分词的匹配思想，这个方法首先将分词问题进行了一定的转化，变成了字的标注问题，即序列标注。该方法通常使用4位标记，规定每个字有4种分类，词首类型B，词中类型M，词尾类型E，单字成词S。当完成对每个字符的标注之后，再通过4种分类完成文本的切分。也就是说，将问题从文本切分转化为了序列标注，而序列标注问题又可以作为分类问题来解决。事实上，基于机器学习的算法非常擅长解决分类问题[11]。常见的统计分词模型有以下几种。

### 2.2.1 隐马尔可夫模型（HMM）

隐马尔可夫模型是一种生成式模型[12]，首先预设了独立性假设，即词与词之间互相独立。此模型可以由一个五元组来描述：观测序列，隐藏序列，隐藏态起始概率，隐藏态之间转换概率（转移概率），隐藏态表现为观测值的概率（发射概率）。从隐藏态初始状态出发，计算下一个隐藏态的概率，并依次计算后面所有的隐藏态转移概率。最终问题就转化为了求解概率最大的隐藏状态序列问题。以观测序列为X，隐状态序列为Y，因果关系为从Y到X。于是有以下公式

$$\mathrm{P}(X,Y)=\prod_{t=1}^{T}P(y_t|y_{t-1})*P(x_t|y_t) \tag{2-1}$$

对于HMM，可以使用动态规划的Viterbi算法进行求解。然而此模型的独立性假设显然是不合理的。由于这个假设，导致无法考虑上下文信息，限制了特征的信息量，最终分词结果也会受到影响。

### 2.2.2 感知机算法

词的感知机算法是一种判别式模型。此模型给每个字符分配一个向量作为特征，再通过构造超平面，将特征空间中的样本分为正负两类，然后使用不断预测答案，计算误差，再利用误差更新参数，优化模型的方法提高预测效果。通过组合设计，再将二分类模型推广到多分类问题，即可应用到中文分词任务。由于每次迭代都会更新模型的所有权重，被误分类的样本会造成很大影响，可以采用平均的方法，在处理完一部分样本后对更新的权重进行平均，称为平均感知机。感知机算法已经能够在中文分词任务上取得非常好的效果了[13]。

### 2.2.3 条件随机场（CRF）

事实上CRF可以看作是HMM的一种扩展求解[14]，不过与HMM不同，CRF是一种判别式模型。该模型同样通过定义条件概率$P(Y|X)$来描述模型，并给定了字符级别的特征。

$$score(l|s)=\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1}) \tag{2-2}$$

简单来说，分词所使用的是Linear-CRF，它由一组特征函数组成，包括权重λ和特征函数$f$，特征函数$f$的输入是整个句子$s$、当前位置$i$、前一个词位$l_{i-1}$，当前词位$l_i$。

总体来说，CRF相比HMM打破了观测独立假设，可以参考更多输入文本的全局特征，并采取了全局归一化，在一定程度上规避了标注偏置的问题。在实验里，CRF还对未登录词具有很好的识别能力。最终，CRF成为了传统中文分词领域最具代表性的算法[15]，缺点为模型复杂，系统开销较大。

# 第 3 章 深度学习方法

最近，神经网络模型迎来了第三次热潮。在历史上，由于缺乏合适的训练算法和高效的硬件资源，神经网络模型曾两次陷入低谷。然而，事物的发展总是曲折的，随着算力的迅速提升和数据的爆炸式增长，以及大规模并行计算的GPU的出现，使得神经网络模型乃至于层数更多的深度神经网络的充分训练成为可能。本世纪初便有人

研究了能否用数值向量表示自然语言词汇，以及如何进行表示的问题[16]。故对于机器学习的中文分词方法而言，特征如何选择，如何对词语进行合理的表示在很大程度上决定了模型的效果，而这也是难点和痛点。相较而言，深度学习的方法可以实现特征的自动抽取和训练，理论上可以更好地保留上下文的信息，而事实上的效果也非常优秀。此类方法的主要缺点在于没有坚实的理论基础，所有研究都建立在实验的基础上，但又无法针对实验结果给出确切的，令人信服的解释。这导致学者们只能通过粗略的理解探索改进的方向，再做大量的实验进行实践性的筛选。近两年，学者们提出神经网络的可解释性亟待研究，并陆续有团队参与了相关工作。Zhang[17]对可解释性领域的工作进行了总结并发表在了IEEE上。

## 3.1 神经网络模型

深度学习领域影响深刻的有卷积神经网络（CNN）[18]，生成对抗网络（GAN）[19]，循环神经网络（RNN），长短期记忆神经网络（LSTM）[20]，transformer[20]等。在NLP领域中最常用的为RNN、LSTM和transformer，他们在各个任务上都有突出的表现。

### 3.1.1 RNN 和 LSTM

这两类神经网络模型非常相似，可以说LSTM是在RNN的基本思想上进行改进的产物。对于普通的神经网络来说，每个神经元节点间的信息难以共享，而RNN在神经元间加上了信息传递，可以使上一个神经元的特征以一定比例传播到当前神经元，我们把它称为记忆功能。这样的结构对于具有序列特性的数据非常有效，而恰好中文分词就可以被转化为序列上的问题。但是RNN也存在许多问题，在进行反向传播时，由于加上了记忆，会更容易出现梯度弥散和梯度爆炸的问题，不易控制。LSTM对RNN的神经元内部进行了改进，添加了多条梯度传播的路径，大大减小了RNN中关于梯度的问题，同时，LSTM还加入了遗忘门，使梯度传播更加稳定可控。总体来说，LSTM能够在一个神经元中保留更多上下文的信息，大大提高了分词准确率，一度取得了中文分词领域的SOTA[22]。

### 3.1.2 transformer

transformer是目前最流行的神经网络模型之一，其抛弃了传统的CNN或RNN结构，整体由self-attention和encoder-decoder组成。

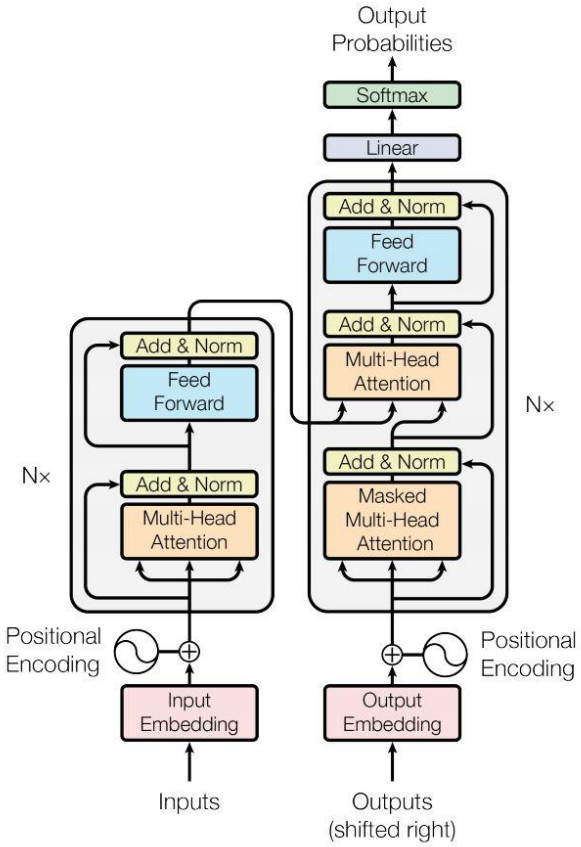


图 3-1 transformer

模型考虑到即使是Bi-LSTM也只能从两种方向分别计算，不能同时考虑序列单位左右的信息，这样一方面，限制了参考的信息量，另一方面，顺序计算时传播的特征量也更容易丢失，因此提出了attention机制。正如论文题目所言，attention机制是整篇文章最重要的东西。它通过计算当前字词与上下文字词的相似度，在训练迭代的过程中捕捉一段文字里字词间的语义特征，成功保存了长距离的有效信息。事实上，attention机制不仅在NLP领域[23]，在整个深度学习相关领域的应用也相当广泛[24]。

## 3.2 预训练模型

前面提到，深度学习的兴起和硬件算力的提高密切相关，而预训练模型就可以把庞大的算力保存在参数中。谷歌在transformer的基础上，提出了BERT模型[25]，通过设置遮蔽语言任务作为预训练目标，第一个步骤学习字词间的语义关系，第二个步骤学习句子间的语义关系。训练完成后，再将模型保存下来，与不同的输出层相结合，仅需继续训练极少数量的参数，即可完成不同的任务需求。BERT是极其有效的，在推

出时便在11项NLP任务中取得了最佳表现。以BERT为例，除了模型的改进，预训练更重要的优点在于可以使用超大数据集和超大算力。应用于各项任务时，仅需使用少量资源即可进行微调，这对广大学术界和工业界都是很有价值的。

# 第 4 章 研究方向展望

解决中文分词的算法模型越来越成熟，但在深度学习不断取得进展的背景下，如表4-1所示，每年都有学者宣布取得新的SOTA。同时，更有研究[26]表明，在中文分词领域，实验中表现出色的模型并不是完美的。下面对近期和未来可能的研究热点进行介绍。

表 4-1 中文分词领域 SOTA

| Model | AS | CITYU | MSRA | PKU |
|---|---|---|---|---|
| Ke et al. (2021) [27] | 97.0 | 98.2 | 98.5 | 96.9 |
| Qiu, Pei, Yan, Huang (2020) [28] | 96.4 | 96.9 | 98.1 | 96.4 |
| Tian, Song, Xia, Zhang, Wang (2020) [29] | 96.6 | 97.9 | 98.4 | 96.5 |
| Meng et al. (2019) [30] | 96.7 | 97.9 | 98.3 | 96.7 |
| Huang et al. (2019) [31] | 96.6 | 97.6 | 97.9 | 96.6 |
| Ma et al. (2018) [22] | 96.2 | 97.2 | 97.4 | 96.1 |
| Yang et al. (2017) [32] | 95.7 | 96.9 | 97.5 | 96.3 |

## 4.1 集成算法

集合机械分词、传统机器学习和深度学习等多种算法，发挥不同算法的优势，规避劣势。比较经典的方法有Huang等[33]提出的利用Bi-LSTM与CRF解决序列标注问题，以及Yao等[34]提出的使用Bi-LSTM-RNN解决中文分词。均已取得较好的效果。

## 4.2 联合模型

传统的流水线式系统容易导致前序任务中错误的传递，从而影响后续任务的效果。故而可以采用多个任务同时训练，使多种语义知识交互便利，相互融合的联合模型。目前已有部分中文分词和词性标注联合模型的相关研究[29,35]。

## 4.3 迁移学习

对于不同领域，不同场景以及不同分词标准下的语料，我们可以尝试捕捉其通用的语义特征，对抗性地去除拟合信息，再使用少量资源对模型进行迁移学习和信息补充，以适用多个专业领域。同时，迁移学习还有望解决特定领域数据集稀少的问题。

## 4.4 多模态

正如CV领域前几天提出的新方法MAE[36]，CV和NLP所使用的模型以及方法正在逐渐趋同，也许未来，我们能够提出一种新模型，合理地将文本和图像的特征表示出来，进行多模态的融合学习。

# 结　论

本文整理归纳了中文分词技术的发展历程以及出现的难点和问题，对各个阶段的代表性算法进行了简要的介绍，在深度学习方法崛起之时给出了一个紧跟潮流的视角，为后来的研究者提供了最新的研究视角。在大数据、人工智能等概念的兴起之下，中文分词也必将顺应时代主流。事实也确实证明，深度学习是目前最为突出、最有希望的方法之一。本文最后指出，在中文分词领域，仍有诸多方向有待探索，如集成算法、联合模型、迁移学习、多模态等，都值得研究人员的深入探究。

# 参考文献

[1] Li X, Ma D, Yin B. Advance research in agricultural text-to-speech: the word segmentation of analytic language and the deep learning-based end-to-end system[J]. Computers and Electronics in Agriculture, 2021, 180: 105908.

[2] 郭风岚,刘辉.中高级水平学习者汉语语序偏误的类型学分析[J].汉语学习,2017(02):98-105.

[3] Emerson T . The Second International Chinese Word Segmentation Bakeoff [C]// Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea. New York, USA: ACL, 2005: 123-133.

[4] 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.

[5] 段慧明,松井久仁於,徐国伟,胡国昕,俞士汶.大规模汉语标注语料库的制作与使用[J].语言文字应用,2000(02):72-77.

[6] 刘江. 大规模汉语语料库分词一致性检验技术研究[D].山西大学,2005.

[7] 周婷,邓慧爱.新词汇来源与发展趋向探究[J].牡丹江大学学报,2018,27(11):97-100.

[8] Grishman R, Sundheim B M. Message understanding conference-6:A brief history[C]//Proceedings of the 16th International Conference on Computational Linguistics-Volume 1, San Mateo, CA, USA, 1996.

[9] 张启宇,朱玲,张雅萍.中文分词算法研究综述[J].情报探索,2008(11):53-56.

[10] 张华平,刘群.基于N-最短路径方法的中文词语粗分模型[J].中文信息学报,2002(05):1-7.

[11] 杨剑锋,乔佩蕊,李永梅,王宁.机器学习分类问题及算法研究综述[J].统计与决策,2019,35(06):36-40.

[12] 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.

[13] Zhang Y, Clark S . Chinese Segmentation with a Word-based Perceptron Algorithm [C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic. USA: ACL, 2007: 840-847.

[14] Lafferty J, McCallum A, Pereira F C N . Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]// Proceedings of the 18th International Conference on Machine Learning, MA, USA. CA, USA: ICMS, 2001: 282-289.

[15] Zhao H, Kit C . Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition [C]// Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing, Hyderabad, India. New York, USA: ACL, 2008: 106-111.

[16] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The journal of machine learning research, 2003, 3: 1137-1155.

[17] Zhang Y, Tiňo P, Leonardis A, et al. A survey on neural network interpretability[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021.

[18] LeCun Y, Bottou L, Bengio Y , et al. Gradient-based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998,86(11):2278-2324.

[19] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[20] Graves A . Long Short-Term Memory[A]// Graves A. Supervised Sequence Labelling with

Recurrent Neural Networks[M]. Berlin: Springer, 2012: 37-45.
[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[22] Ma J, Ganchev K, Weiss D. State-of-the-art Chinese Word Segmentation with Bi-LSTMs[C]//EMNLP. 2018.
[23] Duan S, Zhao H. Attention Is All You Need for Chinese Word Segmentation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 3862-3872.
[24] Chaudhari S, Mithal V, Polatkan G, et al. An attentive survey of attention models[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-32.
[25] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.
[26] Fu J, Liu P, Zhang Q, et al. RethinkCWS: Is Chinese Word Segmentation a Solved Task?[J]. arXiv preprint arXiv:2011.06858, 2020.
[27] Ke Z, Shi L, Sun S, et al. Pre-training with Meta Learning for Chinese Word Segmentation[J]. arXiv preprint arXiv:2010.12272, 2020.
[28] Qiu X, Pei H, Yan H, et al. A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder[J]. arXiv preprint arXiv:1906.12035, 2019.
[29] Tian Y, Song Y, Xia F, et al. Improving Chinese word segmentation with wordhood memory networks[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8274-8285.
[30] Meng Y, Wu W, Wang F, et al. Glyce: Glyph-vectors for chinese character representations[J]. arXiv preprint arXiv:1901.10125, 2019.
[31] Huang W, Cheng X, Chen K, et al. Toward fast and accurate neural chinese word segmentation with multi-criteria learning[J]. arXiv preprint arXiv:1903.04190, 2019.
[32] Yang J, Zhang Y, Dong F. Neural word segmentation with rich pretraining[J]. arXiv preprint arXiv:1704.08960, 2017.
[33] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
[34] Yao Y, Huang Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation[C]//International Conference on Neural Information Processing. Springer, Cham, 2016: 345-353.
[35] 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.
[36] He K, Chen X, Xie S, et al. Masked Autoencoders Are Scalable Vision Learners[J]. arXiv preprint arXiv:2111.06377, 2021.

| 标题 | 作者 | 相似度 | 上传时间 | 论文状态 | 支付状态 | 操作 |
| --- | --- | --- | --- | --- | --- | --- |
| 1120192236-赖昱行-中文分... | | 9.64% | 2021-11-20 | 完成 | 未支付 | 查看 删除 |