整理了近五年中文分词的文章，意图探索研究趋势。

文章来源来自于几大nlp顶会：ACL、EMNLP、COLING、NAACL.

在阅读16、17年文章的时候，发现了引用了这些文章的来自于AAAI、IJCAI的一些文章，也算是一个小Tip。



主要从以下几方面进行总结：

* **论文名：**
* **作者：**
* **机构：**
* **主要工作（解决了什么）：**
* **主要方法（使用了什么）：**
* **数据集：**
* **取得的成果：**
* **缺陷和不足：**
* **下一步工作：**

简要的统计与总结如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 论文数量 | 2014 | 2015 | 2016 | 2017 | 2018 | 总计 |
| ACL | 5 | 3 | 4 | 3 | 0 | 15 |
| EMNLP | 3 | 2 | 0 | 2 | 1 | 8 |
| COLING、NAACL | 1-0 | 0-0 | 0-1 | 0-0 | 1-0 | 3 |
| AAAI、IJCAI | /-/ | /-/ | /-/ | 0-2 | 1-1 | 4+ |
| 总计 | 9 | 5 | 5 | 7 | 4 |  |

ps：“/”代表未进行统计

自己总结的一些趋势（不够专业，仅供参考）：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 2014 | 2015 | 2016 | 2017 | 2018 |
| 无监督 | 2 | 0 | 0 | 0 | 0 |
| 半监督 | 1 | 1 | 0 | 1 | 1 |
| 有监督 | 6 | 4 | 5 | 6 | 3 |
|  |  |  |  |  |  |
| 神经网络 | 1 | 3 | 5 | 6 | 4 |
| 传统方法 | 8 | 2 | 0 | 0 | 0 |
|  |  |  |  |  |  |
| 联合建模 | 0 | 0 | 1 | 1 | 0 |
| 特定领域 | 专利 |  |  | 微文本 | 医学 |
| 重点在于搭建or改进神经网络 | 1/9 | 3/5 | 3/5 | 3/7 | 2/4 |
| 解决领域适配问题 | 3/9 | 1/5 | 0/5 | 2/7 | 3/4 |

2014

**论文名：A Joint Model for Unsupervised Chinese Word Segmentation**

**作者：**Miaohong Chen Baobao Chang Wenzhe Pei

**机构：**教育部计算语言学重点实验室， 北京大学电子工程与计算机科学学院

**主要工作（解决了什么）：**一种无监督中文分词的联合模型，提升了分词的精读

**主要方法（使用了什么）：**结合了基于字符的模型、非参数贝叶斯语言模型和基于良好性的模型的优点。

**数据集：** PKU and MSRA

**取得的成果：**提高了无监督分词模型的精度、较强的中文分词歧义解决能力

**论文名：Effective Document-Level Features for Chinese Patent Word Segmentation**

**作者：**Si Li Nianwen Xue

**机构：**Brandeis大学中文处理组

**主要工作（解决了什么）：**手工分割了大量的专利数据，设计了文档级的特性来捕捉专利中科学和技术术语的分布特征，

**主要方法（使用了什么）：**CRF模型，提取了字符、词语、文档级别的特征

**数据集：**142 Chinese patents（自己标注）

**取得的成果：**解决了专利领域的分词问题

**论文名：Domain Adaptation for CRF-based Chinese Word Segmentation using**

**Free Annotations**

**作者：**Yijia Liu , Yue Zhang , Wanxiang Che , Ting Liu , Fan Wu

**机构：**新加坡理工大学，麻省理工学院社会计算和信息检索研究中心，哈尔滨工业大学

**主要工作（解决了什么）：**通过将各种免费标注源转换为部分标注数据的一致形式。并构造一种可以同时使用全部标注和部分标注数据进行训练的CRF变体，研究了分词的**领域适配问题**。

**数据集：**自由数据+免费标注数据

**取得的成果：**研究了分词的**领域适配问题。**自由数据的有效性，发现它们对于提高分割精度是有用的。

**论文名：Max-Margin Tensor Neural Network for Chinese Word Segmentation**

**作者：**Wenzhe Pei Tao Ge Baobao Chang∗

**机构：**教育部计算语言学重点实验室，北京大学电子工程与计算机科学学院

**主要工作（解决了什么）：**提出了一种新的中文分词神经网络模型——最大边缘张量神经网络(MMTNN)，显式地模拟了标签和上下文字符之间的交互。提出了一种有效地提高模型效率和避免过拟合风险的张量因子分解方法。

**主要方法（使用了什么）：**模型明确地模拟了标签和上下文字符之间的交互，从而获取了更多的语义信息。在神经网络模型中引入张量分解用于序列标记任务，加快了模型的训练和推理，防止了过拟合。

**数据集：**PKU and MSRA datasets

**取得的成果：**我们的模型比以前的神经网络模型具有更好的性能，并且我们的模型可以在最小的特征工程条件下获得具有竞争力的性能。

**下一步工作：**进一步扩展模型，并将其应用于其他结构预测问题。

**论文名：**Automatic Corpus Expansion for Chinese Word Segmentation by Exploiting the Redundancy of Web Information

**作者：**Xipeng Qiu, ChaoChao Huang and Xuanjing Huang

**机构：**上海市智能信息处理重点实验室，复旦大学计算机科学学院

**主要工作（解决了什么）：**处理一个新的领域而没有足够的标注语料库时，监督方法就不能很好地工作。

提出了一种自动扩充外域文本训练语料库的方法

**主要方法（使用了什么）：**通过网络提供的大量相关的容易分割的句子，来分割一个复杂和不确定的分段。挑选出一些可靠的分段句，并将它们添加到语料库中，从而扩充语料库。

**数据集：**both CTB6.0 and CTB7.0 datasets

**取得的成果：**提出了一种自动扩充外域文本训练语料库的方法

**下一步工作：**我们的方法的长期目标是建立一个在线的、不断学习的系统，能够识别困难的任务，并从众包中寻求帮助。

**论文名：**Two Knives Cut Better Than One:Chinese Word Segmentation with Dual Decomposition

**作者：**Mengqiu Wang、Rob Voigt、Christopher D. Manning

**机构：**Stanford University （Computer Science & Linguistics） Department

**主要工作（解决了什么）：**中文分词主要有两种方法:基于单词的分词和基于字符的分词，两种方法各具优势。先前的研究表明，将这两种模型结合起来可以提高分割性能。提出了一种有效地结合两种分割方案的强度的方法，使用一种有效的双分解算法进行联合推理。

**主要方法（使用了什么）：**

**数据集：**SIGHAN 2005 第二届国际汉语分词

**取得的成果：**本文提出了一种基于对偶分解的中文分词方法。可以对现有的CWS系统进行联合解码，比单独的任何一个系统都更加准确和一致，实现了迄今为止在该任务的标准数据集上所报告的最佳性能。

**论文名：**Empirical Study of Unsupervised Chinese Word Segmentation Methods for SMT on Large-scale Corpora

**作者：**Xiaolin Wang、 Masao Utiyama 、Andrew Finch、 Eiichiro Sumita

**机构：**国家信息与通信技术研究所 （日本？）

**主要工作（解决了什么）：**无监督分词(UWS)可以在没有标注数据的情况下为统计机器翻译(SMT)提供领域自适应分词，双语的UWS甚至可以优化分词的对齐。提出了一种有效的统一的基于pyp的单语和双语无监督方法。

**主要方法（使用了什么）：**本研究旨在为统计机器翻译建立一个更好的中文分词模型。该算法利用基于双语特征的对齐自动学习的词界信息，提出了一种较好的分割模型。

**数据集：**Chinese SIGHAN-MSR corpus

**取得的成果：**提高无监督分词模型的精度

**论文名：**Semi-Supervised Chinese Word Segmentation Using Partial-Label Learning With Conditional Random Fields 基于条件随机场的部分标签学习的半监督中文分词方法

**作者：**Fan Yang 、Paul Vozila

**机构：**Nuance Communications Inc. 从事语音识别软件、图像处理软件及输入法软件研发销售的公司

**主要工作（解决了什么）：**Wikipedia数据中的标点符号和实体标记在句子中定义了一些单词边界。采用条件随机场的部分标签学习方法，将这些有价值的知识用于半监督汉语分词。

**数据集：**CTB-6 corpus

**取得的成果：**CTB-6 test set is 95.98% in F-measure

**论文名：**Toward Better Chinese Word Segmentation for SMT via Bilingual Constraints

**作者：**Xiaodong Zeng、 Lidia S. Chao、 Derek F. Wong、 Isabel Trancoso、 Liang Tian

**机构：**澳门大学计算机与信息科学系

**主要工作（解决了什么）：**针对SMT任务提出了一种新的CWS模型。该模型的目的是维护树库数据的语言分割监督，同时集成由bitexts产生的有用的双语分割。

**主要方法（使用了什么）：**从基于字符的对齐中学习词的边界; 将学习到的单词边界编码为GP约束; 在GP约束下，利用PR框架对CRFs模型进行训练。

**数据集：**BLEU、NIST、METEOR 机器翻译质量

**取得的成果：**该模型能较好地实现用于机器翻译的分词

2015

**论文名：**Semi-supervised Chinese Word Segmentation based on Bilingual

Information

**作者：**Wei Chen、Bo Xu

**机构：**自动化所

**主要工作（解决了什么）：**半监督的分词模型，包含三个层次的双语语言特征学习的层叠对数线性模型

**数据集：**1998人民日报语料库、Bakeoff-2 PKU、AS

**取得的成果：**精度超过优于目前最先进的单语和双语半监督方法。

**论文名：**Gated Recursive Neural Network for Chinese Word Segmentation

**作者：**Xinchi Chen, Xipeng Qiu∗ , Chenxi Zhu, Xuanjing Huang

**机构：**复旦大学上海市智能信息处理重点实验室

**主要工作（解决了什么）：**提出了一种用于中文分词的门控递归神经网络(GRNN)。由于GRNN相对较深，我们还采用了一种有监督的分层训练方法来避免梯度扩散问题。

**数据集：**PKU, MSRA and CTB6 NLPCC 2015 dataset微博文本

**取得的成果：**模型优于以往的神经网络模型和最先进的方法。

**下一步工作：**GRNN在其他序列标记任务上的应用。

**论文名：**Long Short-Term Memory Neural Networks for Chinese Word Segmentation

**作者：**Xinchi Chen, Xipeng Qiu∗ , Chenxi Zhu, Pengfei Liu, Xuanjing Huang

**机构：**复旦大学上海市智能信息处理重点实验室

**主要工作（解决了什么）：**提出了一种新的汉语分词神经网络模型，采用长短时记忆(LSTM)神经网络

**数据集：**PKU, MSRA and CTB6

**取得的成果：**模型优于以往的神经网络模型和最先进的方法。

**下一步工作：**改为双向

**论文名：**Synthetic Word Parsing Improves Chinese Word Segmentation

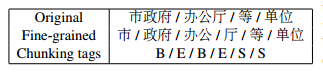
**作者：**Fei Cheng、 Kevin Duh、 Yuji Matsumoto

**机构：**奈良信息科学研究所

**主要工作（解决了什么）：**提出了一种利用人工分词器提高汉语分词性能的新方法，该解析器分析单词的内部结构，并试图将未登录词转换为词汇内的细粒度子单词。我们提出了一个管道CWS系统，该系统首先预测这种细粒度的分割，然后对输出进行分段，以重构原始的分词标准。

**数据集：**PKU和MSR

**取得的成果：**提升了精度。在OOV召回方面有了实质性的改进。



**论文名：**Accurate Linear-Time Chinese Word Segmentation via Embedding Matching

**作者：**Jianqiang Ma 、Erhard Hinrichs

**机构：**德国图宾根大学语言学部

**主要工作（解决了什么）：**本文提出了一种嵌入匹配的中文分词方法，该方法对传统的序列标记框架进行了推广，并利用了分布式表示的优点。在此基础上，提出了一种基于基准语料库的贪心分割算法。

**数据集：**PKU MSR

**取得的成果：**提升了精度

2016

**论文名：**Neural Word Segmentation Learning for Chinese

**作者：**Deng Cai and Hai Zhao∗

**机构：**上交计算机系

**主要工作（解决了什么）：**在本文中，我们提出了一种新的神经网络框架，它彻底消除了上下文窗口，可以利用完整的分割历史。我们的模型采用基于字符的门控组合神经网络生成候选词的分布式表示，然后将其输入长期短期记忆(LSTM)语言评分模型。

**主要方法（使用了什么）：**新的中文分词神经网络框架

**数据集：**PKU、MSR

**取得的成果：**提升精度。神经网络分词模型

**论文名：**An Empirical Study of Automatic Chinese Word Segmentation for Spoken Language Understanding and Named Entity Recognition 汉语自动分词用于口语理解和命名实体识别的实证研究

**作者：**Wencan Luo∗ Fan Yang

**机构：**匹兹堡大学

**主要工作（解决了什么）：**分词通常被认为是许多汉语自然语言处理任务的第一步，但它对这些后续任务的影响却相对缺乏研究。在对新数据应用现有的分词器时，如何解决不匹配问题?一个更好的分词器会产生更好的NLP后续任务性能吗?

**主要方法（使用了什么）：**将分词输出作为附加特征，采用部分学习的自适应技术，利用n-best分词表。

**数据集：CTB6、PKU、NER、**NER 3-fol

**取得的成果：**解决分词应用到新数据不匹配的问题--3个方法。（提高了SLU、NER精度）

**论文名：**A New Psychometric-inspired Evaluation Metric for Chinese Word Segmentation

**作者：**Peng Qian Xipeng Qiu∗ Xuanjing Huang

**机构：**复旦大学上海智能信息处理重点实验室

**主要工作（解决了什么）：**一种新的基于心理测量学的汉语分词评价指标

**主要方法（使用了什么）：**从心理测量学的基本思想入手，在困难与容易、奖励与惩罚之间取得平衡，提高评价的准确性。

**数据集：PKU、MSR、NCC**

**取得的成果：**一种新的基于心理测量学的汉语分词评价指标。实际评价结果表明，所提出的评价指标给出的评价结果更加合理、易于区分，与人的评价结果具有较好的相关性。传统的精度、召回率和F-score。

**论文名：**Dependency-based Gated Recursive Neural Network for Chinese Word Segmentation

**作者：**Jingjing Xu and Xu Sun

**机构：**北京大学计算机语言学教育部重点实验室

**主要工作（解决了什么）：**一种基于依赖的门控递归神经网络

**主要方法（使用了什么）：**为了将局部特征与长距离依赖相结合，提出了一种基于依赖的门控递归神经网络。局部特征首先由双向长短期记忆网络收集，然后通过门控递归神经网络将其组合并细化为长距离依赖。

**数据集：PKU、MSRA、CTB6**

**取得的成果：提高了精度**

**论文名：**Transition-Based Neural Word Segmentation

**作者：**Meishan Zhang1 and Yue Zhang2 and Guohong Fu1

**机构：**黑龙江大学计算机科学与技术学院

**主要工作（解决了什么）：**一种基于词的中文分词神经模型，将人工设计的离散特征替换为基于词的分词框架中的神经特征。

**数据集：**PKU、MSR

**取得的成果：**基于词的汉语分割神经网络模型。该模型可以方便地利用离散特征，从而得到了一个与以往工作相比性能最好的组合模型。提高精度。

2017

**论文名：**Fast and Accurate Neural Word Segmentation for Chinese

**作者：**Deng Cai1,2, Hai Zhao1,2,∗, Zhisong Zhang1,2, Yuan Xin3, Yongjian Wu3, Feiyue Huang3

**机构：**上海交通大学计算机科学与工程系

**主要工作（解决了什么）：**现有神经模型的训练和工作过程都存在计算效率低下的问题。提出了一种具有均衡的字和字符嵌入输入的贪婪神经词分割器，以克服现有的缺陷。

**主要方法（使用了什么）：**一种新型的字符 - 字平衡机制，用于字的表示生成。通过去掉不必要的设计，建立更有效的字符组合模型。最大间隔训练期间的早期更新策略。

**数据集：PKU 95.8、MSR 97.1 （速度快）**

**取得的成果：**更简单、更快、更准确。神经网络模型

**论文名：**Adversarial Multi-Criteria Learning for Chinese Word Segmentation

汉语分词的对抗性多准则学习

**作者：**Xinchi Chen, Zhan Shi, Xipeng Qiu∗ , Xuanjing Huang

**机构：**复旦大学上海智能信息处理重点实验室

**主要工作（解决了什么）：**提出了针对CWS的对抗性多准则学习，将来自多个异构分割准则的共享知识集成在一起。

**数据集：**MSRA AS PKU CTB CKIP CITYU NCC SXU

**取得的成果：**在8个具有异构分割标准的语料库上的实验表明，与单标准学习相比，每个语料库的性能都有了显著的提高。

**论文名：**Multi-Grained Chinese Word Segmentation

**作者：**Chen Gong, Zhenghua Li∗ , Min Zhang, Xinzhou Jiang

**机构：**苏州大学

**主要工作（解决了什么）：**利用三个SWS数据集的注释异构性，构建了用于模型训练和调优的大型伪MWS数据集。然后我们用真正的MWS注释手工注释1500个测试句子。提出了三种将MWS转换为构成解析和序列标记的基准测试方法。

**数据集：CTB、MSR、PPD**

**取得的成果：**提出并解决了多粒度中文分词。

**论文名：**Neural Word Segmentation with Rich Pretraining

**作者：**Jie Yang∗ and Yue Zhang∗ and Fei Dong

**机构：**新加坡理工大学

**主要工作（解决了什么）：**对大型中文文本进行预处理，嵌入字符和单词，可以提高分割准确率

**主要方法（使用了什么）：**预训练是利用外部资源提高精度的一种方法

**数据集：**CTB6 96.2 PKU 96.3 MSR 97.5 AS 95.7 CityU 96.9 Weibo 95.5

**取得的成果：**研究了丰富的外部资源（预训练方法），以加强神经分词。。

IJCAI

**论文名：**Segmenting Chinese Microtext: Joint Informal-Word Detection and Segmentation

with Neural Networks

**作者：**Meishan Zhang, Guohong Fu∗, Nan Yu

**机构：**黑龙江大学

**主要工作（解决了什么）：**微文本中的非正式词汇问题

**主要方法（使用了什么）：**提出了一种通过分词和非正式词检测同时进行的方法来增强中文微文本分割的联合模型。

**数据集：**CTB6.0 and the released Weibo corpus

**取得的成果：**该联合模型可以显著提高微博数据集的分割性能，提高幅度超过3%。

**论文名：**Word-Context Character Embeddings for Chinese Word Segmentation

**作者：**Hao Zhou∗ Zhenting Yu∗ Yue Zhang Shujian Huang Xinyu Dai Jiajun Chen

**机构：**南京大学

**主要工作（解决了什么）：**本文提出了一种基于分词标签信息的文本上下文字符嵌入方法。该方法将标签分布信息打包到嵌入中，可以看作是知识参数化的一种方式。

**数据集：**CTB6 96.2 PKU 96.0 MSR 97.8

**取得的成果：** 提出了半监督神经网络的词-上下文字符嵌入，使分割模型对域内数据更加准确，对域外数据更加稳健。

**论文名：**A Feature-Enriched Neural Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging

**作者：**Xinchi Chen, Xipeng Qiu∗ , Xuanjing Huang

**机构：**复旦大学上海智能信息处理重点实验室

**主要工作（解决了什么）：**本文提出了一种用于汉语分词和词性标注联合作业的特征丰富的神经模型。具体来说，为了模拟传统离散特征模型的特征模板，我们使用不同的滤波器对复杂的混合特征进行卷积和池化建模，然后利用递归层的长距离依赖信息。

**数据集：CTB、PKU、NCC**

**取得的成果：** joint S&T task 词性标注精度提高

2018

2018

**论文名：**State-of-the-art Chinese Word Segmentation with Bi-LSTMs

**作者：**Ji Ma、 Kuzman Ganchev、 David Weiss

**机构：**Google AI Language

**主要工作（解决了什么）：改进模型--**pretrained embeddings、dropout、超参数调整

**数据集：**AS CITYU CTB6 CTB7 MSR PKU UD

98.03 98.22 97.06 97.07 98.48 97.95 97.00

**取得的成果：**为汉语分词的进一步发展提供了两条重要的证据。首先，比较不同的模型体系结构需要仔细地调整和应用最佳实践，以便获得严格的比较。其次，如果不进一步努力收集数据，对神经体系结构的迭代可能不足以解决剩余的类分割错误。

**论文名：**Adaptive Multi-Task Transfer Learning for Chinese Word Segmentation in Medical Text

**作者：**Junjie Xing、 Kenny Q. Zhu、 Shaodian Zhang

**机构：上交**

**主要工作（解决了什么）：**提出了一个自适应多任务转移学习框架和三个不同设置的模型实例。

**数据集：**PKU MSR WEIBO

**取得的成果：**提出了一种新的医学领域中文分词框架。提高了医学领域分词。三个医学数据集的标注工作。

**论文名：**Neural Networks Incorporating Unlabeled and Partially-labeled Data for Cross-domain Chinese Word Segmentation

**作者：**Lujun Zhao, Qi Zhang, Peng Wang, Xiaoyu Liu

**机构：**复旦大学智能信息处理重点实验室

**主要工作（解决了什么）：**

**主要方法（使用了什么）：**提出了一种新的神经网络模型，利用未标记和部分标记的数据来完成跨领域的CWS任务。为了利用无标记数据，通过门机制将两种字符级语言模型与分割模型相结合。通过修改损失函数，利用部分标记数据对模型进行训练。

**数据集：**the labeled People’s Daily (PD) dataset. SIGHANBakeoff 2010 Chinese Wikipedia

**取得的成果：**解决缺乏注释数据的资源贫乏领域的CWS问题

**论文名：**Neural Networks Incorporating Dictionaries for Chinese Word Segmentation

**作者：**Qi Zhang, Xiaoyu Liu, Jinlan Fu

**机构：**复旦大学智能信息处理重点实验室

**主要工作（解决了什么）：**试图解决汉语分词任务中词典与神经网络相结合的问题。研究了基于神经网络的汉语分词任务字典集成问题。针对CWS任务，我们提出了两种将从字典中提取的信息集成到基于神经网络的方法中的方法。

**数据集：**PKU MSR AS CITYU CTB6 2010bakeoff

96.5 97.8 95.9 96.9 96.4

**取得的成果：**合并字典可以显著提高神经分词能力。比目前最先进的神经网络模型和领域适应模型取得了更好的效果。