北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

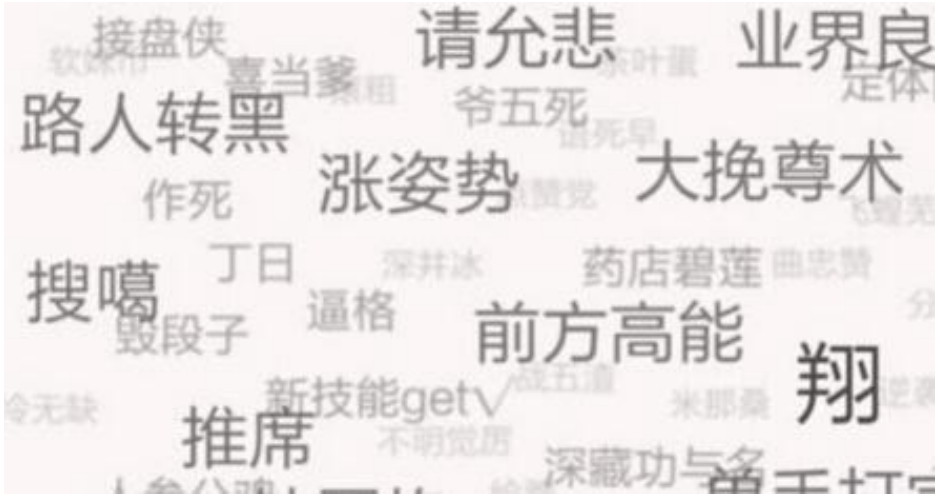
新词发现

汇报人：崔博远、邱家刚、栗怡、闫文麟、赵亚洲、胥玉斌

时　间：2021年10月25日

CONTENTS

随着社会的发展，产生了大量的新词



具体到自然语言处理领域中，有新词和未登录词两种概念。

未登录词指分词处理系统无法识别的词汇或者说已有词库中未出现过词汇。

新词指旧词新用或者未登录词。

新词发现就是发现新词的一种技术。

- 机器翻译

新词的出现使翻译的难度增大，一些新词并不能准确获取其意思，因此新词识别对机器翻译的准确性和翻译效果都有了更高的要求。

- 信息抽取技术

信息抽取的文档通常以互联网为主要信息来源，不断出现的网络新词使得信息抽取的规则和模式需要进行改变。

- 文本情感分析

情感分析中存在大量具有一定情感倾向的新词，研究具有情感色彩的新词对情感分析任务至关重要。

新词的研究基本分为三个阶段。

- 20 世纪 80 年代以前，个别研究者根据出现的少量新词针对个人体验在新词新意现象和新词术语等方面进行了研究。新词的研究处于孕育期。

- 20世纪 80 年代到 21 世纪初，新词的研究进入了增长期。在吕叔湘先生的呼吁下，大部分人开始关注新词并对新词的产生、构成语义等方面进行了研究，关注点涉及新词语词典、教学、社会文化等方面。

- 新世纪至今，新词被更多人接纳。这个时期经济科技迅猛发展，研究者融入了多种研究方法和技术手段，研究视角也不断丰富，新词的研究出现了新的繁荣景象。

- 基于统计的方法通过分析新词的特点，根据词语内部关联程度较高的特征来识别新词。通常以大规模语料库为训练语料，利用有监督的机器学习模型处理新词发现问题，或把它转化为命名实体识别等相关问题并在此基础上进行新词识别。该方法领域通用能力强，移植性好。但具有运算成本大，数据稀疏和准确率低的缺点。
- 基于规则的方法通过总结新词的构词特点建立人工规则，利用规则库筛选新词。利用规则的方法找新词通常针对性强、准确率高。但因为新词产生速度快、词语的构成灵活多变，构建规则库工作量大、成本高、扩展性差。
- 基于统计和规则相结合的方法融合了统计学的特征计算和规则准确率高的优点来提高系统的性能，是近几年主流的方法。
- 基于深度学习的方法：通用性增强，对于低频新词的识别率更好，利用上下文信息的能力更强。

- 对新词来说构词标准灵活多变，对新词的定义也不尽相同，很难找到统一方法进行新词识别工作。
- 由于数据的稀疏性造成新词识别中低频词的识别率偏低，识别难度大。
- 目前很难根据新词出现的时间信息和词语的词形、词义用法的变化发现新词。

评价指标

- 精确率：新词发现结果中正确识别的词数占识别为新词总数的比例。
- 召回率：新词发现结果中正确识别的新词数占实际新词数的比例。
- F数：精确率和召回率的调和值，即为综合考虑的效果。

CONTENTS

特点

不需要机械分词的词典

解决机械分词基本解决不了的出现歧义和新词发现的问题

分类

有监督方法

无监督方法

有

利用标注语料，将新词发现看作分类或序列标注问题：

- 基于文本片段的某些统计量，以此作为特征训练二分类模型
- 基于序列信息进行序列标注直接得到新词，或对得到的新词再进行判定。

通常用 HMM、CRF、SVM 等机器学习算法实现。

无

不依赖于任何已有的词库、分词工具和标注语料，仅根据词的共同特性，利用统计策略将一段大规模语料中可能成词的文本片段全部提取出来，然后再利用语言知识排除不是新词的“无用片段”或者计算相关度，寻找相关度最大的字与字的组合。接下来，再对这些文本片段作一次清洗与过滤，最后，把所抽取得到的词和已有的词库进行比较，就能得到新词，即可加入新词词库。因此，上述过程可简化为两个步骤：

- 构建词库
- 新词比对

基于统计的新词发现

构建词库

词频

凝聚程度

自由程度

口 词频

频次

阈值

凝聚程度

凝聚程度用以衡量相邻字组合成词语的程度，如果两个人经常待在一起，我们往往会认为他们的关系很亲密，而字与字之间的“亲密关系”则通过凝聚程度来表示。

凝聚程度的计算方式：

词组出现的概率除以构成词组的各个词语出现概率的乘积，然后对结果做对数处理（可以理解为对计算得到的概率值进行了一次映射）。

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

以“的电影”和“电影院”为例：

的电影 = 的 + 电影

电影院 = 电影 + 院

电影：0.01，院：0.01，电影院：0.001

p（电影院）/p（电影）p（院）：0.001/(0.01*0.01)=10

电影：0.01，的：0.2，的电影：0.002

p（的电影）/p（的）p（电影）：0.002/(0.01*0.2)=1

上面的结果表明“电影院”更可能是一个有意义的搭配，而“的电影”则更像是“的”和“电影”这两个成分偶然拼到一块。凝聚程度的计算方式很大程度上受到文本切分的影响。

口 自由程度

词语作为汉语中的一个基本语义单元，具备一个显著的特征——可以灵活地应用到不同的场景中。例如“机器学习”，上下文可以搭配很多动词和名词“学习人工智能知识”、“从事人工智能行业”。但对于“人工智”这个词语来说，上文依然可以搭配很多词语，下文却基本上只能搭配“能”。那么，我们可以认为“人工智”不是一个完整的词语。换个角度来看，词语的自由程度可以理解为词语之间的相关性弱，换言之，词语的独立性高。

如何计算自由程度?

也就是如何衡量当前文本片段的上下文可搭配词语的丰富程度呢?

口 信息熵

matrix67提出可以使用信息熵进行度量，其计算公式如下：

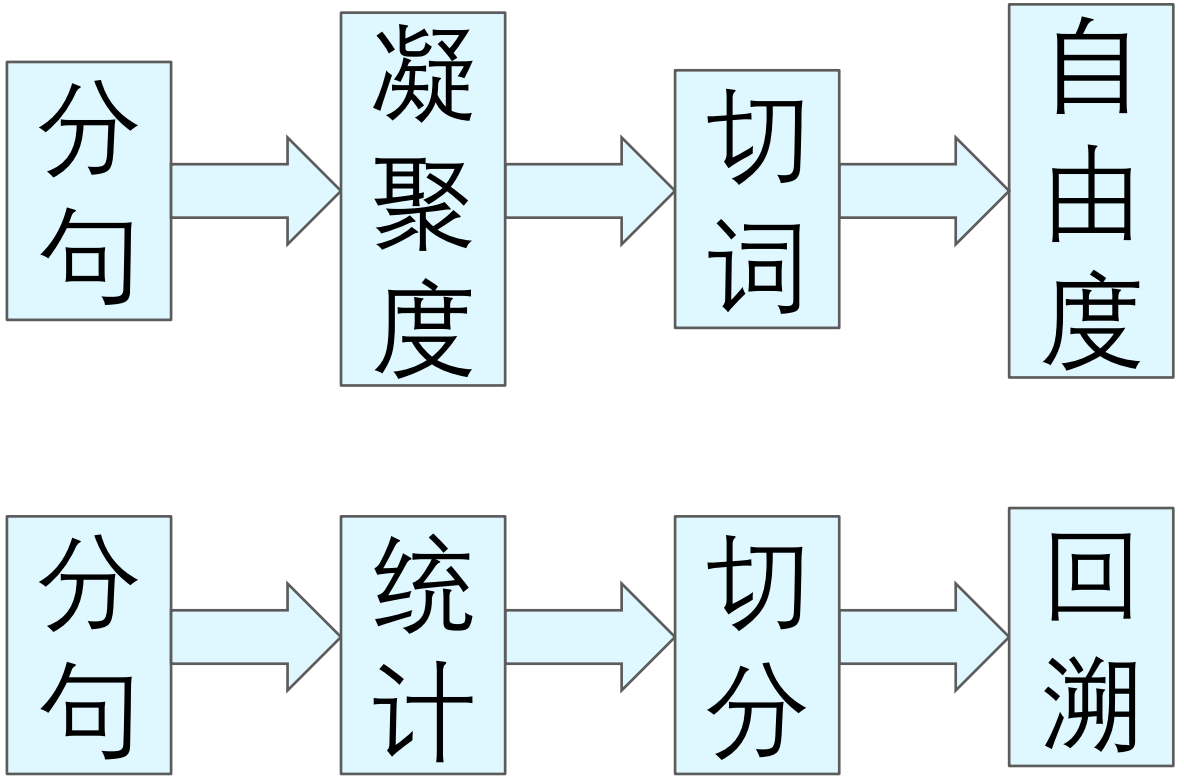
$$H = \sum_{i=1}^{n} P(i) * logP(i)$$

“吃葡萄不吐葡萄皮不吃葡萄倒吐葡萄皮”，“葡萄”一词出现四次，其中左邻字分别为 {吃, 吐, 吃, 吐}，右邻字分别为 {不, 皮, 倒, 皮}。左邻字的信息熵为 – (1/2) · log(1/2) – (1/2) · log(1/2) ≈ 0.693，右邻字的信息熵则为 – (1/2) · log(1/2) – (1/4) · log(1/4) – (1/4) · log(1/4) ≈ 1.04。当前文本片段的上文和下文可搭配词语越丰富，则其左信息熵和右信息熵越大。

基于统计的新词发现


```
['年底', '就在', '显出', '新年', '年的', '接着', '着一', '爆竹', '的大', '没有', '已经', '在这', '这一', '回到',
'我的', '鲁镇', '镇的', '然而', '所以', '只得', '他是', '是我', '四叔', '是一', '一个', '先前', '什么', '么大',
'改变', '单是', '也还', '见面', '之后', '说我', '了说', '知道', '因为', '的还', '还是', '但是', '的了', '于是',
'是不', '不多', '个人', '房里', '第二', '出去', '看了', '几个', '天也', '他们', '们也', '家中', '都在', '祝福',
'一年', '好运', '女人', '人的', '福礼', '了五', '起来', '男人', '自然', '仍然', '然是', '如此', '—', '只要', '
雪花', '大的', '的有', '那么', '成一', '分明', '放在', '还在', '我又', '无聊', '去一', '似乎', '未必', '明天', '
走了', '直到', '的事', '也就', '我不', '那是', '出来', '河边', '见她', '而且', '的眼', '眼睛', '来的', '的我', '
这回', '的人', '人们', '她的', '全不', '脸上', '仿佛', '似的', '的;', '只有', '她是', '她一', '了我', '她来', '
回来', '来了', '了她', '这样', '你是', '得多', '不到', '到她', '她却', '的话', '着就', '就是', '她走', '的说', '
死了', '上也', '时候', '自己', '回答', '呢我', '我在', '想这', '或者', '者不', '不如', '希望', '我想', '说那', '
就有', '地狱', '家的', '唉唉', '不见', '实在', '我也', '不再', '觉得', '不安', '答话', '有些', '大约', '别的', '
意思', '的呢', '发生', '说是', '话的', '即使', '一句', '往往', '说话', '了一', '的雪', '雪天', '现在', '在不', '
虽然', '有我', '以为', '恐怕', '果然', '听到', '但不', '了只', '待到', '们的', '短工', '打听', '不是', '怎么', '
死的', '然的', '去了', '不过', '并不', '渐渐', '终于', '的脸', '脸色', '忽而', '立刻', '他的', '也不', '不很', '
完了', '早已', '坐在', '了的', '还要', '响的', '舒畅', '的她', '有一', '家里', '女工', '头上', '皱眉', '四婶', '
是在', '是她', '模样', '只是', '着眼', '将她', '工钱', '大家', '的也', '几天', '里还', '打柴', '了;', '春天', ';
他', '本来', '力气', '都说', '笑影', '掏米', '看见', '山里', '说她', '去那', '主人', '人家', '阿呀', '淘米', '是
大', '两个', '人来', '住她', '说这', '儿子', '烧火', '了可', '我们', '闹得', '呀我', '来说', '着她', '太太', '总
是', '后来', '去的', '真是', '打算', '里去', '她就', '抬到', '那时', '的小', '角上', '孩子', '眼光', '在她', '只
好', '小篮', '阿毛', '门槛', '去寻', '神气', '的去', '走开', '上的', '笑容', '她说', '一看', '了这', '开去', '许
多', '柳妈', '她便', '鲁镇的', '四老爷', '是一个', '有什么', '第二天', '的—', '四叔的', '祥林嫂', '在鲁镇', '的
人们', '还可以', '的话来', '的时候', '然而她', '起来了', '说不清', '她大约', '了然而', '有别的', '别的事', '一句话
', '的女人', '也没有', '他们的', '出去了', '自己的', '不知道', '四叔家', '是自己', '的婆婆', '小叔子', '个男人', '
忘却了', '了一个', '四叔说', '起来的', '的也有', '的故事', '她的话', '没有什么', '祥林嫂的', '四叔家里', '卫老婆子
', '她的婆婆', '到祥林嫂', '祥林嫂你']
```

改进思路



举例：林心如和易烊千玺支撑着共和国各项目的顺利进展。

一、统计

统计文本片段中2grams、3grams、…、ngrams，计算其内部凝固度，设置阈值，构建集合G。举例：林心如、易烊千玺、共和、共和国、支撑、撑着、各项、项目、顺利、进展

二、切分

结合上述grams对语料进行切分（粗糙的分词），并统计频率。举例：林心如，和，易烊千玺，支撑着，共和国，各项目，的，顺利，进展。

三、回溯

如果它是一个小于等于n字的词，那么检测它在不在G中，不在就出局；如果它是一个大于n字的词，那个检测它每个n字片段是不是在G中，只要有一个片段不在，就出局。举例：林心如，和，易烊千玺，共和国，的，顺利，进展。

基于规则的发现方法是比较传统的方法，主要是指通过匹配所制定的规则来发现新词。一般是通过语言学专家根据词汇学原理来创造匹配模板，然后通过程序进行新词匹配。

目前，基于规则的新词发现方法大致可以分为两类：

- 通过总结新词的构词规则，构建新词规则库，再通过匹配相应规则来发现新词
- 构建新词过滤规则库，将一些明显不符合构词法的词组过滤掉。

一、构建新词规则库的方法

由于需要对新词的构词结构进行观察和总结，所以需要对汉语的构词法和词的结构规律有一定的了解，这对于新词语的识别十分重要。新词的构词规则一般有两种：

◘ 常规的构词规则

大部分的新词的词性结构依旧遵循基本构词法的规则，包括：
“名词+名词” 、“名词+形容词”、“名词+动词” 、“名词+量词”
“形容词+形容词”、“形容词+名词”、“形容词+动词”、“动词+动词”
“动词+名词” 、“动词+形容词” 等10种。

例如：打工人 = 打工（动词） + 人（名词）
永远的神 = 永远的（形容词） + 神（名词）
气氛组 = 气氛（名词） + 组（名词）

特殊的构词规则

新词的组成成份除一部分遵循常规的构词原则外，介词、方位词、语气助词等都被赋予了新的构词能力。

例如：水吧、书吧= 水、书（名词）+ 吧（语气助词）

在线 = 在（介词） + 线（名词）

了解构词规则对于建立一个准确有效的规则库十分重要，建立规则库之后通过正则表达式对规则进行表示，从而实现对新词的抽取。

二、构建新词过滤规则库的方法

构建新词过滤规则库的方法就是根据构词法的有关原则,将候选词中存在的明显的不成词成份去除。

例如：设A、B、C、D代表四个任意汉字

若A为副词，　　B为其它词性，且A位于句首，则A B被过滤掉；

若A为其它词性，B为副词，　　且B位于句尾，则A B被过滤掉；

若A为其它词性，B为助词，　　且B位于句尾，则A B被过滤掉；

若AB、ABC、ABCD中存在连词，则将其过滤。

以上示例为目前一些常用的过滤规则，这种方法不需要总结新词的构词规则。

三、优势与局限性

因为基于规则的新词发现方法要进行规则匹配，所以规则的制定是前提条件，这样虽然可以对某些特定情况下的新词会有很好的效果，但同时因为规则的局限性，会出现新词发现不全，遗漏等问题，同时，制定并维护规则也比较费时费力，所以这种方法的适用性和移植性较差。

基于规则和基于统计相结合的新词发现

一、相关技术和工具

深度学习

机器学习领域中一个较新的研究方向，而且近几年来更是被广泛使用在各个领域中，非常流行。深度学习可以提取数据之间的规律，它所需的数据可以有多种类型，包括文字、图像和声音等等。它的最终效果就是可以让机器识别人类世界中的各种信号。

LSTM网络

Long Short Term 网络一般就叫做 LSTM，是一种 RNN 特殊的类型，可以学习长期依赖信息，通过刻意的设计来避免长期依赖问题。

CRF层

对于序列标注任务，考虑相邻标签之间的关系对于选择最佳的标签链是很有必要的。

例如，在词性标注任务中，形容词后面不可以跟动词，副词后面不可以跟名词。在传统的 Softmax 层中，这样的规律关系是无法发现的，就可能产生一些错误的预测。但是用 CRF 层替代后就可以避免上述错误，因为它可以发现标签之间的一个组合规律。

深度学习框架

深度学习框架是进行模型训练的工具，能够使模型构建更便捷。因为框架封装了很多常用结构的代码，例如 LSTM 节点或 CNN 节点，原本需要一大段代码来定义的元素在使用框架的情况下只需要一行就可以调用，因此框架可以自动构建出所需要的模型。

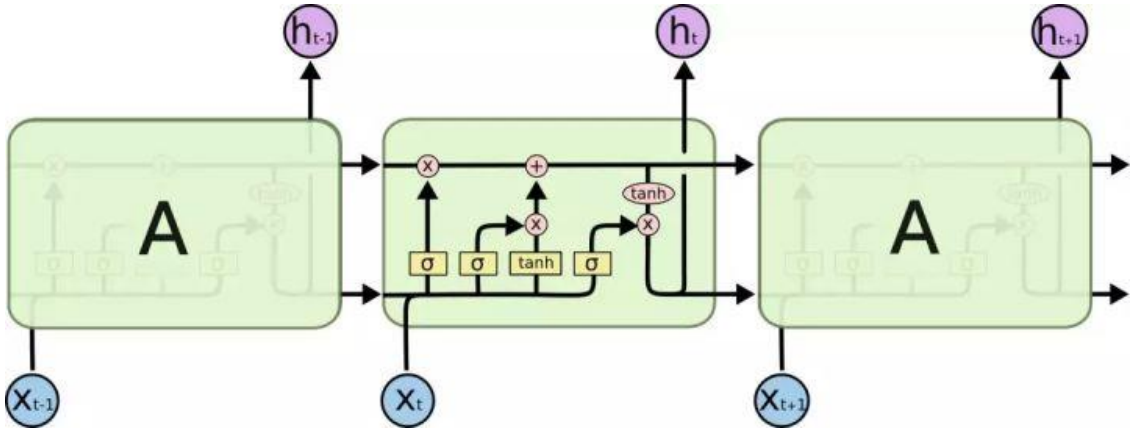
RPC调用

RPC（Remote Procedure Call Protocol）是远程调用的一种协议。它是用于客户端与服务端之间的调用的。但是它的好处在于客户端在调用服务器端的时候不需要知道服务端的实现细节，甚至不需要知道调用方式，就可以实现整个调用过程，就像在调用本地方法一样简单。

二、特征介绍

- **词性**：因为要研究的是将多个旧词拼凑为一个新词的问题，所以能拼凑成新词的每一个旧词的词性可以作为一个考虑因素。
- **词长**：拼凑为新词的旧词的词语长度也是一个需要考虑的因素。如果每一个旧词的词长之和过长，那它合成新词的可能性就要低一些。
- **上下文信息熵**：实质是词语之间的自由度。如果一个词语的上下文信息熵越大，那它与别的词经常在一起出现的机会就越小，经常是独立出现在文本中，这样与其它旧词合并为新词的可能性就越小，反之就越大。
- **词语间凝固度**：词语间凝固度与上下文信息熵恰恰相反，说的是词语相互联结的程度。如果两个词语经常在一起出现，那么就说明这两个词语的词语间凝固度较大，那这些词语的组合就较有可能形成一个新词。

三、研究方法



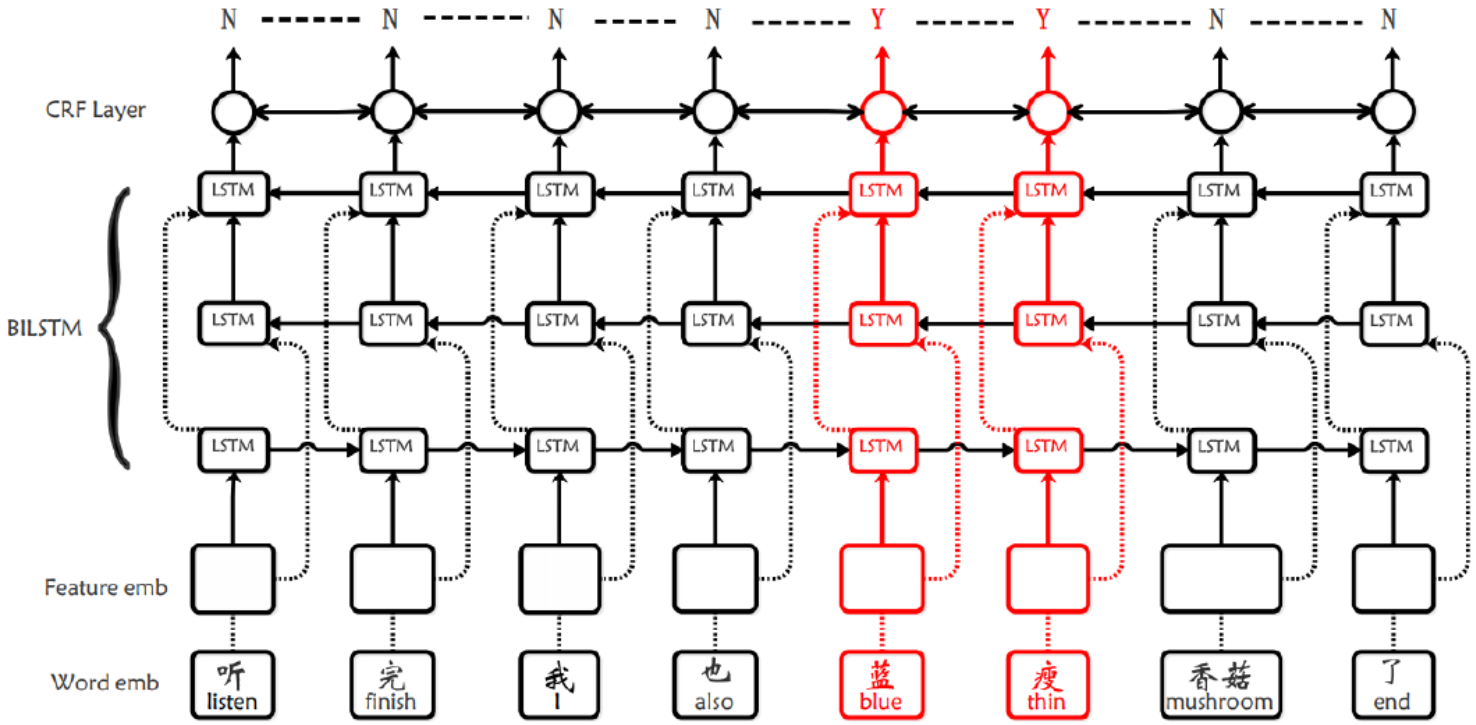
LSTM总体架构

RNN

- 隐藏层单元
- 只使用正向序列信息

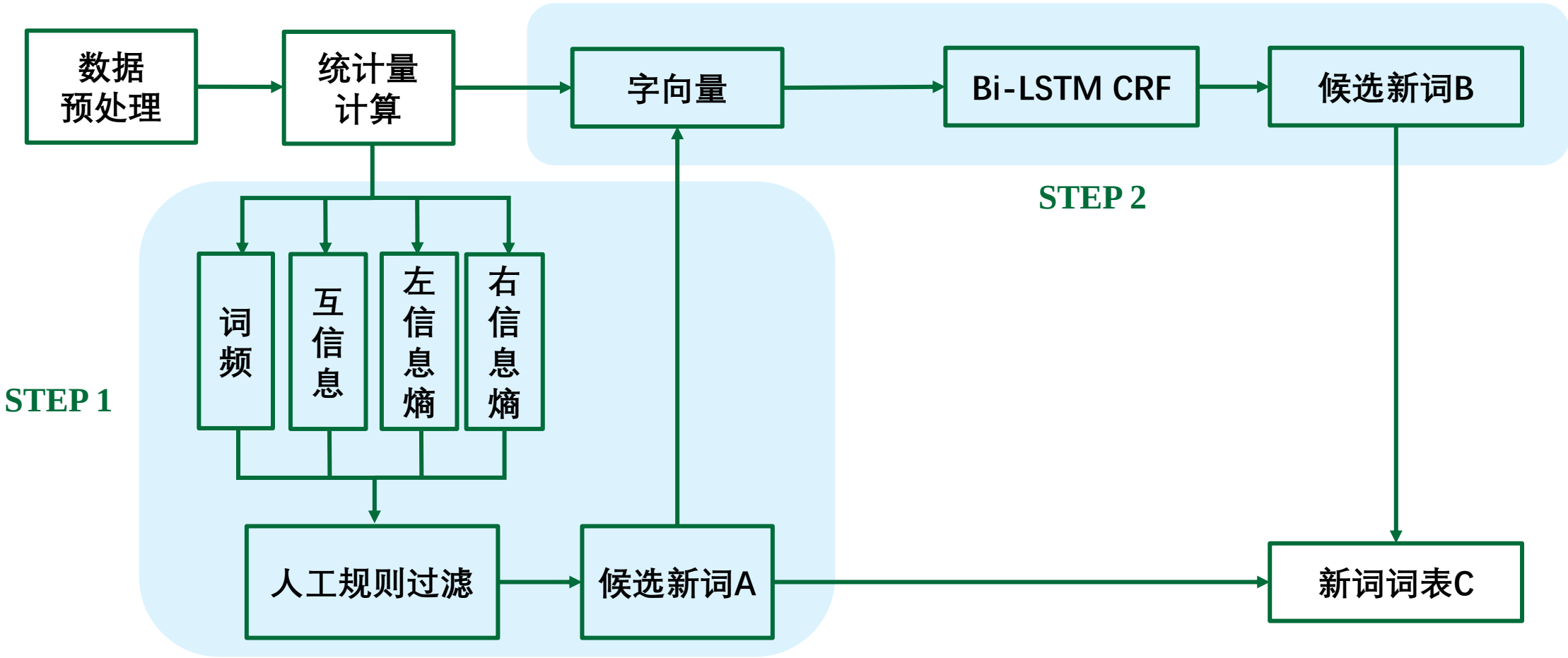
Bi LSTM

- 输入门，输出门，遗忘门
- 充分利用上下文信息



Bi-LSTM+CRF 模型在新词发现中的应用

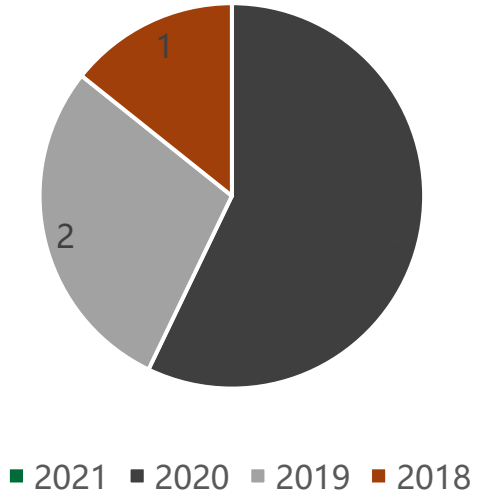
基于深度学习的新词发现



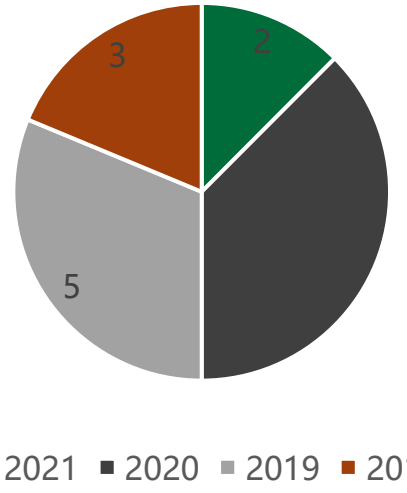
CONTENTS

dblp检索

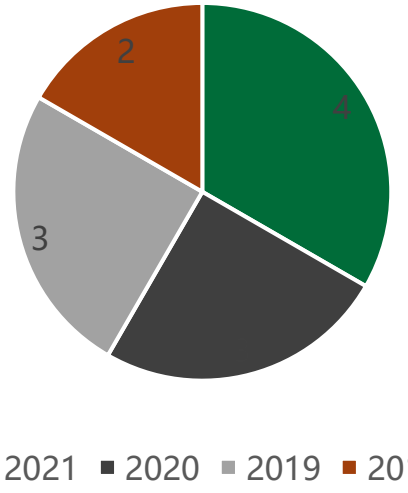
new word discovery

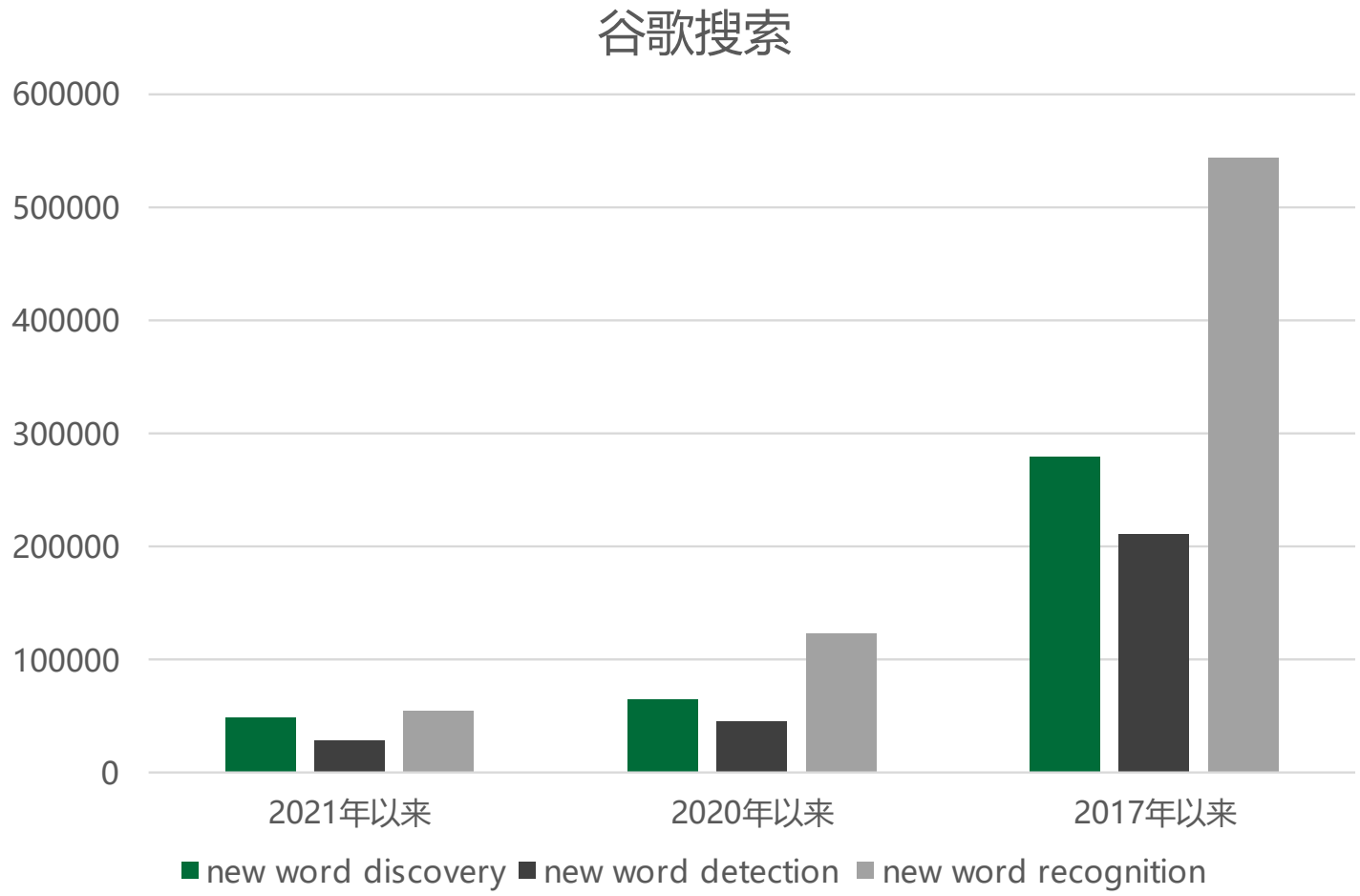


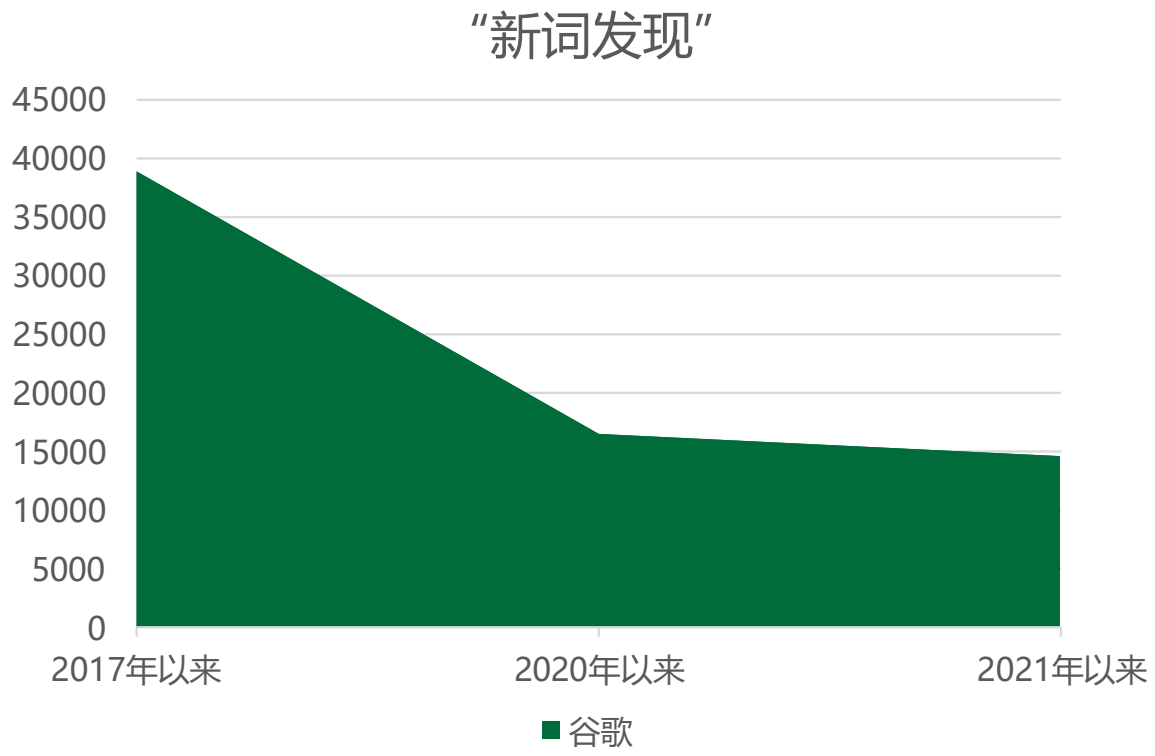
new word detection



new word recognition




谷歌搜索
600000
500000
400000
300000
200000
100000
0
2021年以来
2020年以来
2017年以来
new word discovery
new word detection
new word recognition



2021 年 10 月使用知网，以“新词发现” 为主题词检索相关文献，共获得291篇文献的标题、关键字和摘要。

一、将中文分词和新词发现结合起来

- *F. Peng, F. Feng, and A. Mc Callum, "Chinese segmentation and new word detection using conditional random fields," Proceedings of the 20th international conference on Computational Linguistics, p.562, Association for Computational Linguistics, 2004.*

使用CRF模型进行分词任务，将置信度高但不被认为是词语的词作为新词加入词典。这些词可以进一步提高分词的精度。

- *X. Sun, H. Wang, and W. Li, "Fast online training with frequency adaptive learning rates for chinese word segmentation and new word detection," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp.253–262, Association for Computational Linguistics, 2012.*

可同时用于中文分词和新词发现的模型，在模型中添加了高维的新特征，并使用在线梯度下降的方法提高速度。

二、使用专家总结的语法规则和相关知识来匹配新词

- *Z. Guo Dong, "A chunking strategy towards unknown word detection in chinese word segmentation," International Conference on Natural Language Processing, pp.530–541, Springer, 2005.*

根据词语原子的构词法将一个或多个词语原子组合到一起来检测位置词语的模型。

- *M. Huang, B. Ye, Y. Wang, H. Chen, J. Cheng, and X. Zhu, "New word detection for sentiment analysis," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.531–541, 2014.*

它通过将一些种子词语放入字典中来总结这些种子词的词法模板，然后用该模板通过匹配文本来提取新词，迭代这个过程，将所提取的新词加入到字典中。

然而，这些方法都需要人工花费大量的时间和精力，给任务的完成带来开销。

三、将用户行为数据考虑在内，以检测新词

- *Y. Zheng, Z. Liu, M. Sun, L. Ru, and Y. Zhang, "Incorporating user behaviors in new word detection," Twenty-First International Joint Conference on Artificial Intelligence, 2009. Zheng*

试图发现特定领域中经常使用一些术语的潜在专家，然后进一步从这些专家的术语中提取领域新词。

四、使用统计特征来发现新词

- *Y. Liang, P. Yin, and S.M. Yiu, "New word detection and tagging on chinese twitter stream," International Conference on Big Data Analytics and Knowledge Discovery, pp.310–321, Springer, 2015.*

利用改进的点互信息(PMI)结合一些基本规则来识别网络新词的无监督方法。

- *W. Li, K. Guo, Y. Shi, L. Zhu, and Y. Zheng, "Improved new word detection method used in tourism field," Procedia Computer Science, vol.108, pp.1251–1260, 2017.*

利用增强互信息(EMI)算法对旅游领域的候选新词进行了过滤。与 EMI 算法相比，PMI 算法只考虑了 2-gram，有一定的局限性，而 EMI 可以考虑 n-gram 的词语关系，从而更准确地提取新词。

- *Chen and M. Sun, "Domain-specific new words detection in chinese," Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017), pp.44–53, 2017.*

提出了一种基于领域词字典模型(D-WDM)的领域新词提取方法，该方法采用了领域词字典模型(D-WDM)来代替传统的 WDM 模型，并利用一个领域得分函数来区分一个词语是来自于普通词字典还是领域词字典。

- *Li, Xia, Bin Wu, and Bailing Zhang. "Unknown Word Detection in Song Poetry." 2016 IEEE First International Conference on Data Science in Cyberspace (DSC). IEEE, 2016.*

使用 SVM 来提取新词。

- *刘昱彤, et al. "基于古汉语语料的新词发现方法." 中文信息学报 33.1(2019): 46-55.*

通过使用 LSTM 网络来提取古汉语中的新词。

- *金宇杰,袁明. 基于 TF-IDF 算法的新词发现系统原理与实现[J]. 信息化研究, 2020,v.46;No.292(05):43-48.*

使用 TF-IDF 算法结合关键词抽取方式来提取新词的方法。

- *Li, Peng, Yongxing Guang, and Tianling Qiao. "Research on Chinese New Word Recognition Method." Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering. 2020.*

基于改进的 PMI 算法和熵相结合的无监督 OOV 识别方法来完成新词发现任务。

传统统计模型的新词识别方法

- HMM
- CRF

基于深度神经网络的新词发现方法

- RNN
- CNN
- transformer

Bert (Bidirectional Encoder Representations from Transformers) 是一种基于深度双向 Transformer 的预训练模型。

优点1：由于使用了 self－attention 完全代替递归与卷积，Bert 能够快速地进行并行计算，从而可以训练更大的模型来对句子的上下文语境进行深度双向表征。

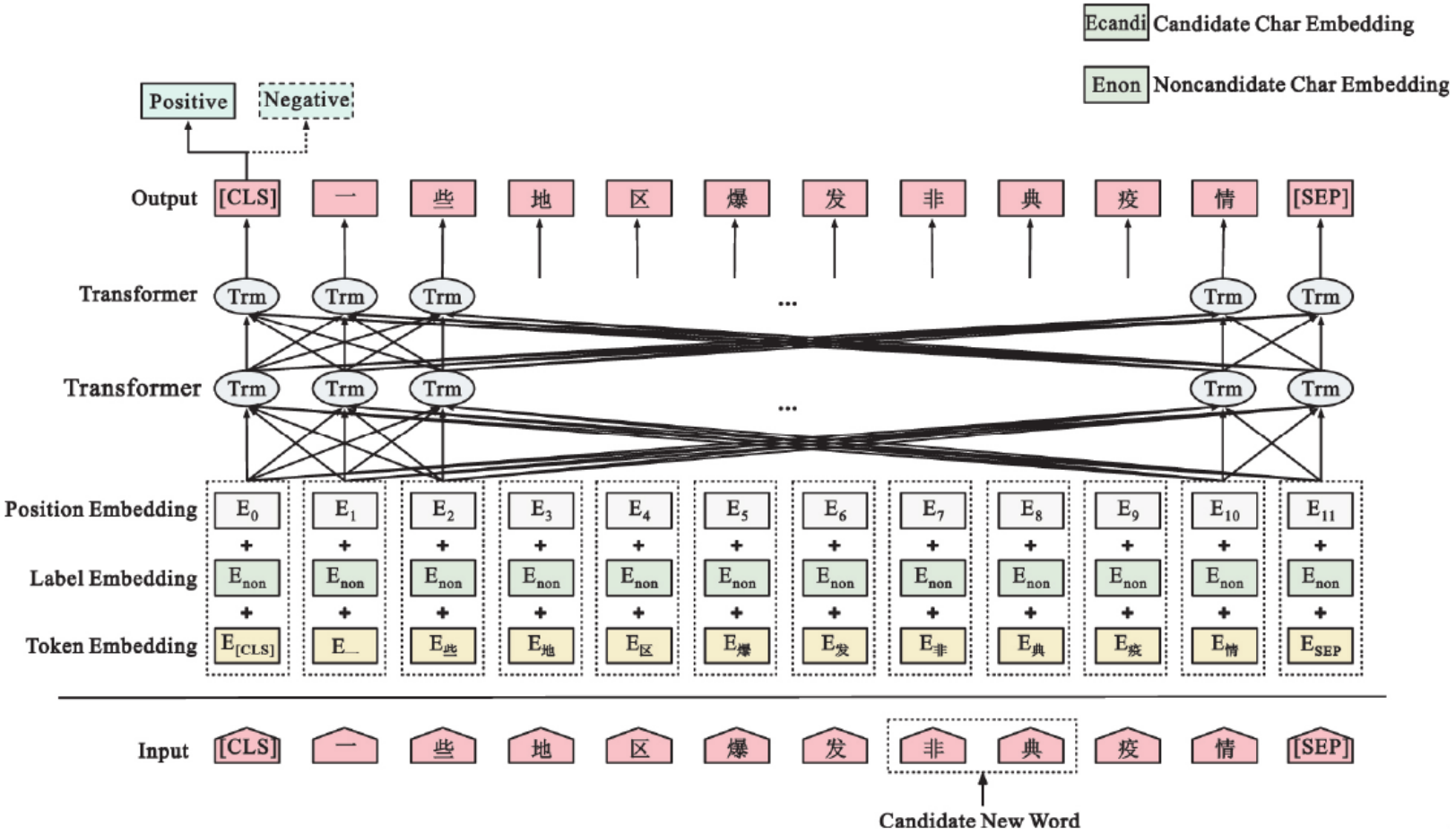
优点2：支持基于 fine－tuning 迁移学习，因此 Bert模型可仅用一个额外的输出层应对各种特定自然语言处理任务，不需对原来的模型作大量修改，仅改变输入和增加输出层即能完成目标。

一、候选新词识别模型

Sentence ={ Char0, Char1, …, Charn}

$$f(x \mid Sentence) = \begin{cases} 1, x \text{ is a valid word} \\ 2, x \text{ is not a valid word} \end{cases}$$

X ={ Chari, Chari+1, …, Chari+l}



基于bert新词识别算法模型结构

二、主动学习策略

主动学习算法抽象模型： $A=(C,U,L,Q,S)$

C 为分类算法， L 为已标注样本，U 为未标记样本

Q 为主动抽样策略，S 为人工标注者

算法过程: 首先从U 中随机标记少量样本加入L 作为初始训练样本集，C 使用L 训练分类模型，同时Q 根据分类模型从U 中筛选出一定数量未标注数据，由S 进行人工标注，并将这部分已标注数据加入L，C 再次使用L 训练分类模型，迭代上述过程直到达到停止条件。

三、基于规则的候选新词过滤

热度规则

突发性规则

合成性规则

CONTENTS

1. 概述　　崔博远
2. 新词发现方法 邱家刚、栗　怡、闫文麟
3. 前沿研究　　赵亚洲
4. Demo展示与讲解　胥玉斌

- 这一部分以GitHub上的Chinese_segment_augment为参考。其基于互信息和左右熵
- 算法过程如下:
  - 使用jieba对语句进行分词
  - 使用字典树存储单词和统计词频
  - 利用信息熵进行新词发现
  - 利用trie树计算左右熵
  - 得出得分score=PMI+min(左熵，右熵)，根据分数由高到低进行排序排序

方法解释:

- 使用词典树对存储分词
- 利用trie树计算互信息PMI
- 利用trie树计算左右熵
- 计算得分score=PMI+min(左熵，右熵)
- 取得得分最高的TOP N作为新词，我们这里选择取得前6个为新词。如果前面待选词属于后面待选词的一部分则删除后面待选的词

  比如:[花呗，蚂蚁花呗] --> [花呗]
- 之后将新词添加到语料库中

原本语料库内容展示

文件内容

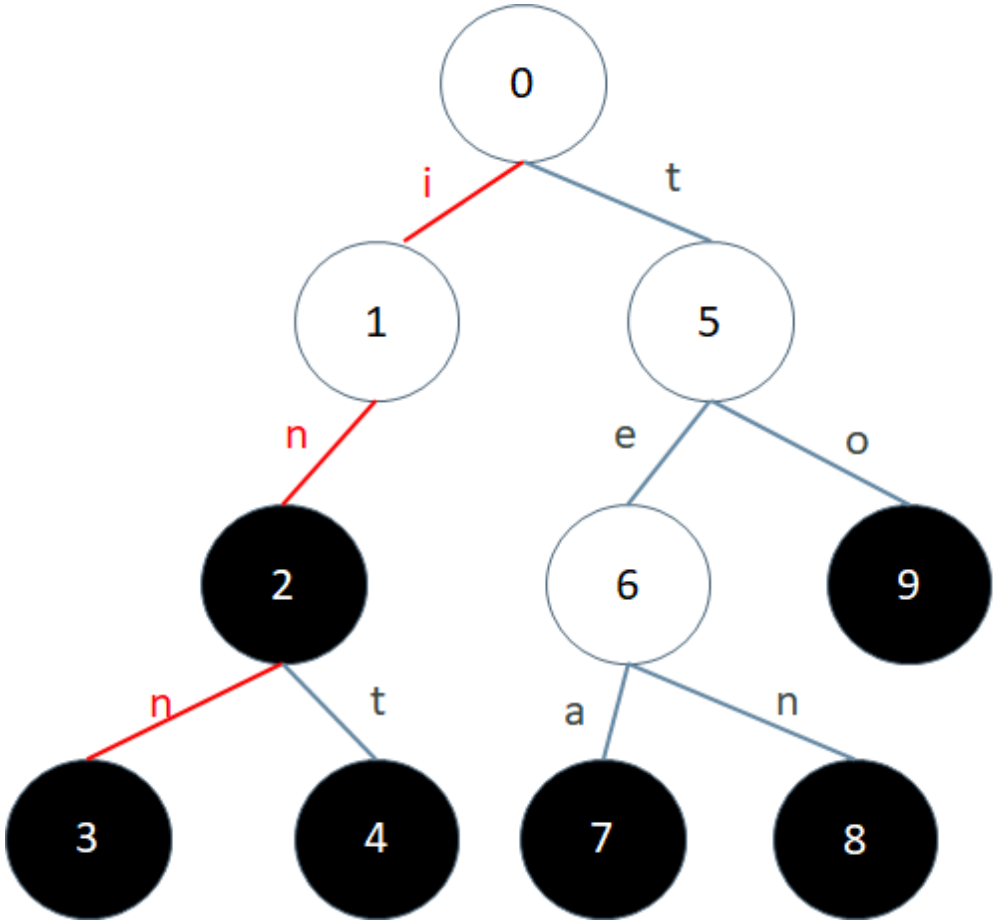
```
AT&T 3 nz
B超 3 n
c# 3 nz
C# 3 nz
c++ 3 nz
C++ 3 nz
T恤 4 n
A座 3 n
A股 3 n
A型 3 n
A轮 3 n
AA制 3 n
AB型 3 n
B座 3 n
B股 3 n
B型 3 n
B超 3 n
B轮 3 n
BB机 3 n
BP机 3 n
C盘 3 n
```

算法结果示例：

输入句子：蔡英文在昨天应民进党当局的邀请，准备和陈时中一道前往世界卫生大会，和谈有关九二共识问题

```
增加了 6 个新词，词语和得分分别为：

##################################
世界卫生大会 ----> 0.4380419441616299
蔡英文 ----> 0.28882968751888893
民进党当局 ----> 0.2247420989996931
陈时中 ----> 0.15996145099751344
九二共识 ----> 0.14723726297223602
世卫大会 ----> 0.13287009686368884
##################################
添加前：
蔡/ 英文/ 在/ 昨天/ 应/ 民进党/ 当局/ 邀请/ 准备/ 和/ 陈时/ 中/ 一道/ 前往/ 世界卫生/ 大会/ 和谈/ 有关/ 九二/ 共识/ 问题/
添加后：
蔡英文/ 在/ 昨天/ 应/ 民进党当局/ 邀请/ 准备/ 和/ 陈时中/ 一道/ 前往/ 世界卫生大会/ 和谈/ 有关/ 九二共识/ 问题/
```



trie树

4.2 结果展示


新词发现
D:/python/大数据/data/demo.txt
选择文件
topN: 6
确定
显示文件内容
台湾“中时电子报”26日报道称，蔡英文今日一早会见“世卫行动团”，她称，台湾虽然
无法参加WHA(世界卫生大会)，但“还是要有贡献”。于是，她
抗埃博拉病毒。
对于台湾为何不能，蔡英文又一次惯性“甩锅”，宣称“中国对
用其极”。
随后，蔡英文还在“世卫行动团”面前自诩，台湾是个“模范生
蔡英文自有一套“道理”：世卫组织“不应该把台湾排除在WHO
有目共睹，然而WHO秘书长却因为政治因素，把台湾这个模范生
艰难，只会成为台湾的动力，台湾用公卫医疗的成就，向世界证
重要的角色”。
随后，蔡英文也不忘感谢在世卫大会上替台“发声”的几个“友
到‘邦交国’的支持和伸援，我们感谢这些国家支持，也感谢(
门的努力”。
不过环环想提醒一句，此次大会上，确实有多个台湾“友邦”受
“邀请台湾作为观察员参加WHA”，然而结果是，立即被大会否
中国国家卫生健康委员会主任马晓伟20日曾对此表示，2009年至2016年，台湾地区连续8年
以“中华台北”名义和观察员身份参加了世界卫生大会。这是在两岸均坚持体现一个中国
原则的“九二共识”基础上，通过两岸协商做出的特殊安排。由于民进党当局迄不承认体
现一个中国原则的“九二共识”，破坏了台湾地区参加世界卫生大会的政治基础。今年台
湾地区收不到参会邀请，责任完全在民进党当局。
而对于蔡英文此次颇费心机出此招数，希望再在WHA上刷一波存在感，岛内网友先看不下去
了，有网友直指“别白忙活了，忙正事吧”，还有网友笑道“典型的狂躁症又发作了”，
也有网友表示蔡英文这一次“又是花纳税人的钱，真是惨！”
台湾卫福部门负责人陈时中
新加入的词
世界卫生大会 ----> 0.4380419441616299
蔡英文 ----> 0.28882968751888893
民进党当局 ----> 0.2247420989996931
陈时中 ----> 0.15996145099751344
九二共识 ----> 0.14723726297223602
世卫大会 ----> 0.13287009686368884
世界卫生_大会--->0.4380419441616299
WHA_上刷--->0.06422422359931641
电子报_26--->0.06357062123824737
马晓伟_20--->0.06357062123824737
痛批_台当局--->0.062649420449710 11
WHA_世界卫生--->0.05857861336196027
三脚_仔是--->0.05792501100089122
尽干_蠢事--->0.05741803762311585
心机_出此--->0.056547894279763436
皇民化_本岛人--->0.05599998356571572
刊发_陈时--->0.05446611775238177

新词发现

D:/python/大数据/data/demo.txt 选择文件 topN: 6 确定 显示文件内容

台湾“中时电子报”26日报道称，蔡英文今日一早会见“世卫行动团”，她称，台湾虽然无法参加WHA(世界卫生大会)，但“还是要有贡献”。于是，她表示要捐100万美元给WHO对抗埃博拉病毒。

对于台湾为何不能，蔡英文又一次惯性“甩锅”，宣称“中国对台湾的外交打压已无所不用其极”。

随后，蔡英文还在“世卫行动团”面前自诩，台湾是个“模范生”。对于台湾无法参会，蔡英文自有一套“道理”：世卫组织“不应该把台湾排除在WHO之外，台湾的健保在世界是有目共睹，然而WHO秘书长却因为政治因素，把台湾这个模范生排除在外”“但外在情势的艰难，只会成为台湾的动力，台湾用公卫医疗的成就，向世界证明，我们可以在世界扮演重要的角色”。

随后，蔡英文也不忘感谢在世卫大会上替台“发声”的几个“友邦”：“这次WHO，台湾得到‘邦交国’的支持和伸援，我们感谢这些国家支持，也感谢(台湾)‘卫福部’和相关部门的努力”。

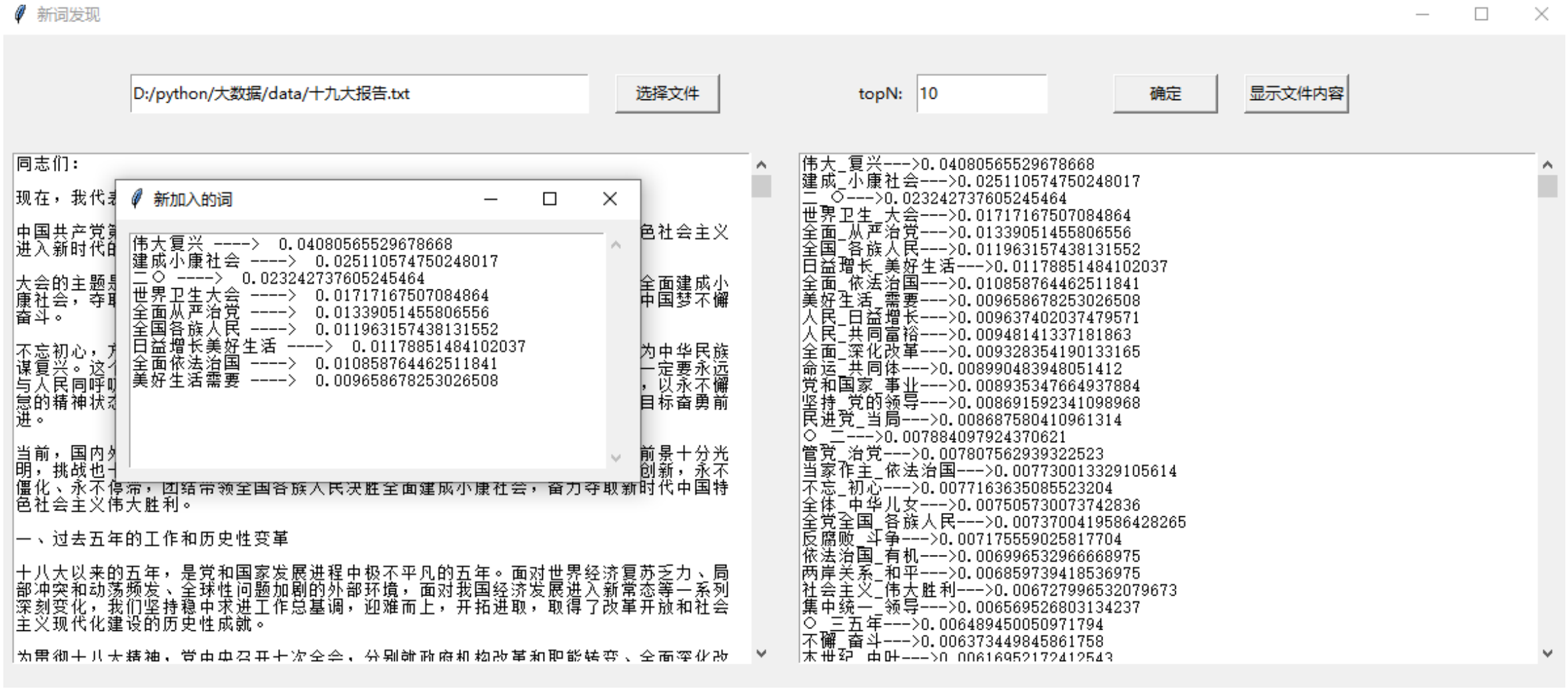
不过环环想提醒一句，此次大会上，确实有多个台湾“友邦”受台当局邀请，向大会提案“邀请台湾作为观察员参加WHA”，然而结果是，立即被大会否决……

中国国家卫生健康委员会主任马晓伟20日曾对此表示，2009年至2016年，台湾地区连续8年以“中华台北”名义和观察员身份参加了世界卫生大会。这是在两岸均坚持体现一个中国原则的“九二共识”基础上，通过两岸协商做出的特殊安排。由于民进党当局迄不承认体现一个中国原则的“九二共识”，破坏了台湾地区参加世界卫生大会的政治基础。今年台湾地区收不到参会邀请，责任完全在民进党当局。

而对于蔡英文此次颇费心机出此招数，希望再在WHA上刷一波存在感，岛内网友先看不下去了，有网友直指“别白忙活了，忙正事吧”，还有网友笑道“典型的狂躁症又发作了”，也有网友表示蔡英文这一次“又是花纳税人的钱，真是惨！”

台湾卫福部门负责人陈时中

世界卫生_大会--->0.4380419441616299
蔡_英文--->0.28882968751888893
民进党_当局--->0.2247420989996931
陈时_中--->0.15996145099751344
九二_共识--->0.14723726297223602
世卫_大会--->0.13287009686368884
原则_九二--->0.12825338342641823
参加_世界卫生--->0.12230763069693351
观察员_身份--->0.11931921337860599
参加_WHA--->0.08914483778518563
届_世界卫生--->0.0856774634835872
参与_世界卫生--->0.0850277688327621
世卫_行动--->0.08204992380820153
国台办_发言人--->0.07348267171991339
台湾地区_参加--->0.0714607122919094
看不下去_有--->0.0704015359044936
上刷_一波--->0.06672022753745997
丢光_尽干--->0.06672022753745997
随后_蔡--->0.06573080212399796
WHA_上刷--->0.06422422359931641
电子报_26--->0.06357062123824737
马晓伟_20--->0.06357062123824737
痛批_台当局--->0.06264942044971011
WHA_世界卫生--->0.05857861336196027
三脚_仔是--->0.05792501100089122
尽干_蠢事--->0.05741803762311585
心机_出此--->0.056547894279763436
皇民化_本岛人--->0.05599998356571572
刊发_陈时--->0.05446611775238177
直指_别白--->0.05405189034161989


新词发现
D:/python/大数据/data/十九大报告.txt
选择文件
topN:
10
确定
显示文件内容
同志们：
现在，我代表
中国共产党第
进入新时代的
大会的主题是
康社会，夺取
奋斗。
不忘初心，方
谋复兴。这个
与人民同呼吸
怠的精神状态
进。
当前，国内外
明，挑战也十
僵化、永不停滞，团结带领全国各族人民决胜全面建成小康社会，奋力夺取新时代中国特
色社会主义伟大胜利。
一、过去五年的工作和历史性变革
十八大以来的五年，是党和国家发展进程中极不平凡的五年。面对世界经济复苏乏力、局
部冲突和动荡频发、全球性问题加剧的外部环境，面对我国经济发展进入新常态等一系列
深刻变化，我们坚持稳中求进工作总基调，迎难而上，开拓进取，取得了改革开放和社会
主义现代化建设的历史性成就。
为贯彻十八大精神，党中央召开十次全会，分别就政府机构改革和职能转变、全面深化改
色社会主义
全面建成小
中国梦不懈
为中华民族
一定要永远
，以永不懈
目标奋勇前
前景十分光
创新，永不
新加入的词
伟大复兴 ----> 0.04080565529678668
建成小康社会 ----> 0.025110574750248017
二〇 ----> 0.023242737605245464
世界卫生大会 ----> 0.01717167507084864
全面从严治党 ----> 0.01339051455806556
全国各族人民 ----> 0.011963157438131552
日益增长美好生活 ----> 0.01178851484102037
全面依法治国 ----> 0.010858764462511841
美好生活需要 ----> 0.009658678253026508
伟大_复兴--->0.04080565529678668
建成_小康社会--->0.025110574750248017
二_〇--->0.023242737605245464
世界卫生_大会--->0.01717167507084864
全面_从严治党--->0.01339051455806556
全国_各族人民--->0.011963157438131552
日益增长_美好生活--->0.01178851484102037
全面_依法治国--->0.010858764462511841
美好生活_需要--->0.009658678253026508
人民_日益增长--->0.009637402037479571
人民_共同富裕--->0.00948141337181863
全面_深化改革--->0.009328354190133165
命运_共同体--->0.008990483948051412
党和国家_事业--->0.008935347664937884
坚持_党的领导--->0.008691592341098968
民进党_当局--->0.008687580410961314
〇_二--->0.007884097924370621
管党_治党--->0.007807562939322523
当家作主_依法治国--->0.007730013329105614
不忘_初心--->0.007716363508552320
全体_中华儿女--->0.007505730073742836
全党全国_各族人民--->0.0073700419586428265
反腐败_斗争--->0.007175559025817704
依法治国_有机--->0.006996532966668975
两岸关系_和平--->0.006859739418536975
社会主义_伟大胜利--->0.006727996532079673
集中统一_领导--->0.006569526803134237
〇_三五年--->0.006489450050971794
不懈_奋斗--->0.006373449845861758

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

谢谢

汇报人：崔博远、邱家刚、栗怡、闫文麟、赵亚洲、胥玉斌
指导老师：张华平
时　间：2021年10月25日