



百度一下

PaddleNLP工具集·全景介绍

龙老师

百度自然语言处理部



百度NLP：扎根产业实践，不断探索前行

服务百度内部产品

支持百度搜索、信息流、
凤巢、贴吧、百科、好看
视频、智能云、文库等
90+ 百度内部产品线

云端服务外部客户

服务**2万+**开发者、**数
千家**企业、日均调用量
1亿+。平台覆盖了金
融、媒体、房产等领域，
积累了大量知名企业客户。

工具集开源

工具集框架
预训练模型
大规模数据集
前沿研究
逐步开放、开源
推动NLP技术的产业化应用

PaddleNLP 开源中文 NLP 工具与数据集



教育版 + AI Studio

工业版

研究版

应用任务

机器翻译

对话模型工具箱

标题生成

中文词法分析

短文本语义匹配

文本情感分类

对话情绪识别

预训练模型 + PaddleHub

前沿研究

V-NET

MRQA

STACL

JTPE

ARNOR

MPM

DuConv

MMPMS

SEEDS

基础网络

序列标注组网集

文本分类组网集

语义匹配组网集

语言生成组网集

语义理解模型: BERT、ERNIE、ELMo

语言模型

大规模数据集

阅读理解数据集

知识驱动对话数据集

语音翻译数据集

信息抽取数据集

实体链指数数据集

飞桨 PaddlePaddle

典型深度学习 学习流程



典型飞桨 代码结构



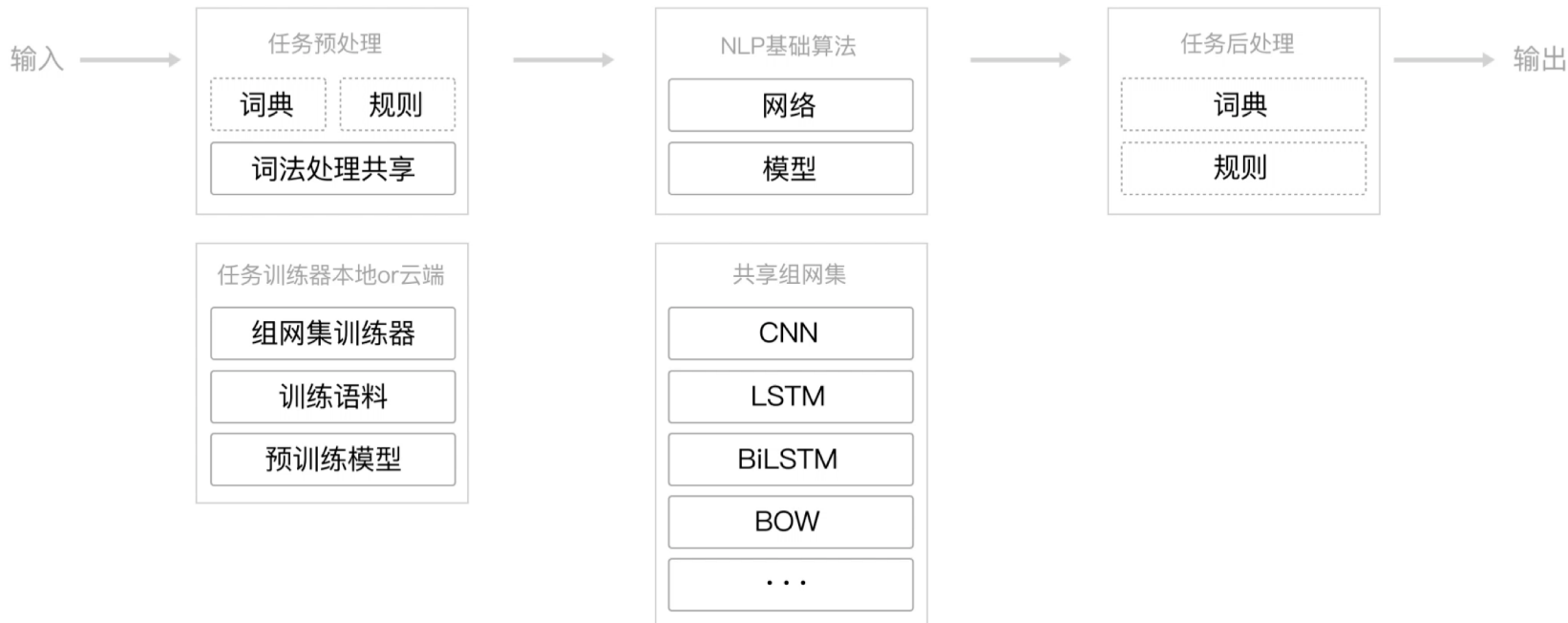
PaddleNLP 目录结构



github.com/PaddlePaddle/models/tree/develop/PaddleNLP

```
├── Research # 百度NLP在research方面的工作集合
├── dialogue_model_toolkit # 对话模型工具箱
├── emotion_detection # 对话情绪识别
├── language_model # 语言模型
├── language_representations_kit # 语言表示工具箱
├── lexical_analysis # 词法分析
├── models # 共享网络
│   ├── __init__.py
│   ├── classification
│   ├── dialogue_model_toolkit
│   ├── language_model
│   ├── matching
│   ├── neural_machine_translation
│   ├── reading_comprehension
│   ├── representation
│   ├── sequence_labeling
│   └── transformer_encoder.py
├── neural_machine_translation # 机器翻译
├── preprocess # 共享文本预处理工具
│   ├── __init__.py
│   ├── ernie
│   ├── padding.py
│   └── tokenizer
├── reading_comprehension # 阅读理解
├── sentiment_classification # 文本情感分析
└── similarity_net # 短文本语义匹配
```


算法网络可灵活拔插替换



网络=函数

models/classification/nets.py

```
def cnn_net(data,
            label,
            dict_dim,
            emb_dim=128,
            hid_dim=128,
            hid_dim2=96,
            class_dim=2,
            win_size=3,
            is_infer=False):
    """
    Conv net
    """
    # embedding layer
    emb = fluid.layers.embedding(input=data, size=[dict_dim, emb_dim])

    # convolution layer
    conv_3 = fluid.nets.sequence_conv_pool(
        input=emb,
        num_filters=hid_dim,
        filter_size=win_size,
        act="tanh",
        pool_type="max")

    # full connect layer
    fc_1 = fluid.layers.fc(input=[conv_3], size=hid_dim2)
    # softmax layer
    prediction = fluid.layers.fc(input=[fc_1], size=class_dim, act="softmax")
    if is_infer:
        return prediction
    cost = fluid.layers.cross_entropy(input=prediction, label=label)
    avg_cost = fluid.layers.mean(x=cost)
    acc = fluid.layers.accuracy(input=prediction, label=label)

    return avg_cost, prediction
```

在工业实践中积淀多年的中文NLP应用任务与模型



工业级预训练模型与数据集

ERNIE 92.0%

BERT 90.2%

BiGRU-CRF 88.0%

中文词法分析

ERNIE 95.4%

BERT 94.3%

BiLSTM 91.8%

文本情感分类

ERNIE 80.6%

BERT 78.6%

TextCNN 76.8%

对话情绪识别

DuReader 30 万真实问题
150 万全网文档

SQuAD 10 万众包问题
2 万维基文档

阅读理解

人工标注12万轮
真实对话
融合360万图谱
知识
兼顾闲聊与任务
对话特色

知识驱动对话

人工标注41万
三元组
20 万中文语句
50 类谓词

信息抽取

语义表示 (ERNIE)

工业级数据集

百度百亿级大数据

`run.sh` 脚本修改如下:

```
# 在train()函数中, 增加--init_checkpoint选项; 修改--vocab_path
--init_checkpoint ./models/textcnn
--vocab_path ./data/vocab.txt
```

`config.json` 配置修改如下:

```
# vocab_size为词典大小, 对应上面./data/vocab.txt
"vocab_size": 240465
```

如果用户基于开源的 ERNIE模型进行Finetune, 需要更新 `run_ernie.sh` 脚本

```
# 在train()函数中, 修改--init_checkpoint选项
--init_checkpoint ./models/ernie_finetune/params
```



3 分钟上手NLP工具集

以文本情感分类为例

paddlepaddle.org.cn/modelanddataset

欢迎沟通交流！



QQ技术交流群
欢迎扫码入群



PaddleNLP官网

<https://nlp.baidu.com/homepage/nlptools/>