NLP
NATURAL LANGUAGE PROCESSING
Natural Language Processing

Applications

Machine Translation
Dialogue System
Deep Question Answering
Intelligent Content Creation
Document Intelligence

Language Understanding
Semantic Analysis
Syntactic Analysis
Lexical Analysis

Document Understanding
Information Extraction
Sentiment Analysis

Language Generation
Response Generation
News Generation
Summarization

Comment Generation
Question Generation
Poetry Generation

Cross-Modality
Cross-Modal Understanding
Cross-Modal Generation

Large-Scale Knowledge Enhanced Pre-trained Models
Open and Domain Knowledge Graph
## ERNIE Knowledge Enhanced Pre-trained Models and Platform for Language Understanding and Generation

<table>
<thead>
<tr>
<th>Application</th>
<th>Web/Video Search</th>
<th>News Feed</th>
<th>Smart Speaker</th>
<th>Document Intelligence</th>
<th>Intelligent Customer Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>EasyDL-NLP (Zero threshold AI Development Platform)</td>
<td>BML-NLP (End-to-end AI Development Platform)</td>
<td>ERNIE-Kit (End-to-end Development Toolkits)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Family</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-Modal</td>
<td>ERNIE-VIL (Vision-Language Understanding)</td>
<td>ERNIE-UNIIMO (Unified模ality)</td>
<td>ERNIE-ViLG (Vision-Language Generation)</td>
<td>ERNIE-SAT (Speech-Language)</td>
<td></td>
</tr>
<tr>
<td>Graph Learning</td>
<td>ERNIE-Sage (Text Graph Computing)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domain-specific Model</td>
<td>ERNIE-Health (Biomedicine)</td>
<td>ERNIE-Law (Law)</td>
<td>ERNIE-Finance (Finance)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task-specific Model</td>
<td>ERNIE-Gen (Language Generation)</td>
<td>ERNIE-Sim (Text Similarity)</td>
<td>ERNIE-iE (Information Extraction)</td>
<td>ERNIE-Search (Document Retrieval)</td>
<td></td>
</tr>
<tr>
<td>Language Understanding &amp; Generation</td>
<td>ERNIE 3.0 (10 Billion Parameters)</td>
<td>PCL-BAIDU Wenxin (260 Billion Parameters)</td>
<td>ERNIE 3.0 Zeus (Task Knowledge Enhanced Model)</td>
<td>ERNIE-M (Cross-lingual model)</td>
<td></td>
</tr>
<tr>
<td>Cross-lingual Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### PCL-BAIDU WENXIN: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation

- **Fine-tuning**
- **Zero-shot Learning**
- **Few-shot Learning**

![Diagram of ERNIE's Knowledge enhanced Continual Pre-training Framework](https://github.com/PaddlePaddle/ERNIE)

- Large Scale
  - 260 billion parameter Pre-trained language model
- Powerful
  - New SoTA on over 60 NLP tasks
- Cost-Effective
  - Parameter Compression Rate of 99.98%

https://github.com/PaddlePaddle/ERNIE
PaddlePaddle Graph Learning Framework
Winner of KDD CUP 2021 / TextGraph 2020 Graph Learning Challenges

PaddlePaddle Reinforcement Learning Framework
Winner of NeurIPS 2018/2019/2020 Learning To Run Challenges

High quality reproduction of over 10+ mainstream RL methods
Easy implementation in CPU, GPU and clusters
Flexible framework, facilitating easy and low-cost secondary development

Applications
- Recommendation
- Search
- Map
- Security
- FinTech
- Biomedicine

Training & Deployment
- Message Passing Across GPUs
- Multi-process Memory Optimization
- Supporting Dynamic Graph & Static Graph Mode

Models
- Walk-based Network Embedding
- Message-Passing Based Graph Neural Network
- Knowledge Graph Embedding

Graph Engines
- CPU Graph Engine
- GPU Graph Engine
- Graph Partition
- Graph Storage
- Graph Sampling
- Graph Random Walk
- Heterogeneous Graph
- Large-scale Parameter Server

Infrastructure
- PaddlePaddle

PARL
PaddlePaddle Reinforcement Learning Framework

https://github.com/PaddlePaddle/PGL

https://github.com/paddlepaddle/PARL
Unified Framework

Unified-Modal Pre-training
Support both language and visual understanding and generation with a unified-modal transformer pre-trained on images, texts and image-text pairs.

Unified Fine-tuning
Support various NLU tasks through language prompts, and achieve good few-shot and zero-shot performances.

Application

Cross-Modal Tagging
Wave Summit 2021 for Deep Learning Developer

Information Extraction
Medical Record
Three days ago, the patient began to have fever and cough with yellow phlegm. The body temperature reached 39.1°C. The blood test and chest x-ray show severe pneumonia.

Disease: severe pneumonia
Manifestation
Temperature: 39.1°C
cough
Examination
X-ray

Text-to-Image Generation
Sunlight shines through the trees lining a gravel path at sunset

Fireworks light up in the city

Entity: Baidu, PaddlePaddle
Event: Wave Summit 2021
Topic: Deep Learning, AI
DUREADER Chinese Question Answering Datasets from Baidu Search

DUREADER 2.0 The largest Chinese MRC dataset containing 300K questions and 1.5M documents
[2018, 2019 Language and Intelligence Competition]
[DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications, In MRQA at ACL 2018]

DUREADER robust A Chinese dataset for evaluating the robustness of MRC models
[2020 Language and Intelligence Competition]

DUREADER yesno A Chinese dataset for evaluating MRC models on yes-no questions
[2020 Chinese Artificial Intelligence Competition]

DUREADER checklist A Chinese dataset for evaluating MRC models from multiple fine-grained aspects
[2021 Language and Intelligence Competition]

DUREADER retrieval The first large scale chinese benchmark for passage retrieval from Baidu Search
[2022 Language and Intelligence Competition]
[DuReaderretrieval: A Large-scale Chinese Benchmark for Passage Retrieval from Web Search Engine, arxiv 2022]

DUREADER vis The first Chinese dataset for open-domain document visual question answering

DuQM A Chinese dataset for evaluating the robustness of question matching models
[2021 CCF Big Data & Computing Intelligence Contest]
[DuQM: A Chinese Dataset of Linguistically Perturbed Natural Questions for Evaluating the Robustness of Question Matching Models, arxiv 2022]

ROCKETQA The Dense Retrieval Models for Open-domain Question Answering

RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain QA, In NAACL2021

RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking, In EMNLP2021

All codes, models and tools are available at github.com/paddlepaddle/rocketqa
- State-of-the-art
- First-Chinese-model
- Easy-to-use
### Products & Services

<table>
<thead>
<tr>
<th>Products &amp; Services</th>
<th>Document Translation</th>
<th>Webpage Translation</th>
<th>Video Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>APP</td>
<td>API</td>
<td>SDK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PC-web</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PC-client</td>
</tr>
</tbody>
</table>

### Models & Algorithms

<table>
<thead>
<tr>
<th>Models &amp; Algorithms</th>
<th>Multi-language</th>
<th>Multi-domain</th>
<th>Cross-modal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Multi-task Learning</td>
<td>Pretraining</td>
<td>Simultaneous Translation</td>
</tr>
<tr>
<td></td>
<td>Multi-agent Learning</td>
<td>Fine-tuning</td>
<td>ASR Error Correction</td>
</tr>
<tr>
<td></td>
<td>Data Augmentation</td>
<td>Adaptation</td>
<td>Interactive Decoding</td>
</tr>
<tr>
<td></td>
<td>Language Clustering</td>
<td>Ensemble</td>
<td>OCR Translation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models &amp; Algorithms</th>
<th>Joint Training</th>
<th>Adversarial Training</th>
<th>Knowledge Distillation</th>
<th>Fast Decoding</th>
</tr>
</thead>
</table>

### Products

<table>
<thead>
<tr>
<th>Product</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Languages</td>
<td>200+</td>
</tr>
<tr>
<td>Developers</td>
<td>500K+</td>
</tr>
<tr>
<td>Characters/Day</td>
<td>100B+</td>
</tr>
</tbody>
</table>

- **PC-client**
- **Simultaneous Translation**
- **Document Translation**
PLATO: Large Scale Diversified Open-Domain Dialogue Generation

Task-Oriented Intent Classification with Cross-Domain Knowledge

Ecosystem of Technologies, Products and Services
# CLOUD API & OPEN-SOURCE LIBRARY

## Cloud API
Programmatic Interfaces to Baidu AI Platform services of NLP

<table>
<thead>
<tr>
<th>Basic services</th>
<th>Application services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Analysis</td>
<td>Document Classification and Tagging</td>
</tr>
<tr>
<td>Dependency Parsing</td>
<td>Summarization</td>
</tr>
<tr>
<td>Semantic Similarity</td>
<td>Text Correction</td>
</tr>
<tr>
<td>Language Model</td>
<td>Sentiment Analysis</td>
</tr>
<tr>
<td>Word Embedding</td>
<td>Emotion Recognition</td>
</tr>
</tbody>
</table>

https://ai.baidu.com

## PaddleNLP
An Easy-to-use NLP Library to Support Developing Wide-range of NLP Applications

### Pre-built NLP Tasks

<table>
<thead>
<tr>
<th>Information Extraction</th>
<th>Sentiment Analysis</th>
<th>Machine Translation</th>
<th>Machine Reading Comprehension</th>
<th>Knowledge Driven Dialogue</th>
<th>Text to SQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Analysis</td>
<td>Syntactic Analysis</td>
<td>Language Model</td>
<td>Text Classification</td>
<td>Text Matching</td>
<td>Text Generation</td>
</tr>
</tbody>
</table>

### Core APIs

#### Transformers

- Auto-encoder
- Auto-regressive
- Seq-to-Seq
- Cross-modal
- Retrieval-based

#### Dataset

- DuReader
- BSTC
- DuConv
- DuIE
- DuEE
- DuEL

PaddlePaddle

https://github.com/PaddlePaddle/PaddleNLP
OPEN DATASETS AND SHARED TASKS

LUGE  An Open-sourced Project of Chinese NLP Datasets

Initiated by Baidu, CCF and CIPSC, with the purpose of accelerating the advancement of NLP techniques, LUGE evaluates NLP models in terms of robustness and adaptability across multiple tasks and multiple domains. Now it has collected 11 tasks, consisting of 37 Chinese datasets by authors from 15 organizations.

NLP Tasks

<table>
<thead>
<tr>
<th>Sentiment Analysis</th>
<th>Machine Reading Comprehension</th>
<th>Information Extraction</th>
<th>Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Similarity</td>
<td>Semantic Parser</td>
<td>Simultaneous Translation</td>
<td>Entity Linking</td>
</tr>
</tbody>
</table>

https://www.luge.ai

LIC  The Most Popular Chinese NLP Shared Tasks

Tasks

- Machine Reading Comprehension: DuReader checklist
- Multi-skill Dialog: LUGE-dialog
- Multi-format Information Extraction: DuIE2.0, DuEE1.0, DuEE-fin

Participants

- 3,512 registered teams and 4,308 individuals
- 49% teams from academia
- 51% teams from industry and individuals
- 10,995 submissions

LIC · 2022  is on going

http://lic2022.cipsc.org.cn
Baidu NLP team is an innovative research and development team working on frontier topics in natural language processing, including semantic representation, understanding and generation. We also dedicate to NLP algorithms in internet applications and industrial applications. We hope to advance the field by publishing our findings, releasing our source codes and organizing shared tasks.

DESCRIPTION

As a member in this team, you will contribute to:

- Keep up-to-date on machine learning and NLP, define and explore frontier topics in NLP.
- Design NLP solutions for Baidu’s products and services.
- Leverage deep learning algorithms in NLP tasks including semantic representation, discourse understanding, sentiment analysis, text generation, question answering, dialogue system, and machine translation.

MINIMUM QUALIFICATIONS

- Master’s or PhD degree in Computer Science or a related field.
- Expertise in python, c/c++.
- Strong knowledge of data structures and algorithms.

PREFERRED QUALIFICATIONS

- Experience in pre-training, few-shot learning, dialogue system, language understanding and generation.
- Publications at leading conferences such as ACL, EMNLP, COLING, or similar.
- Self-motivated team player with excellent communication skills and strong sense of responsibility.

Location: Beijing/Shenzhen, China

Please reach out to nlp-job@baidu.com