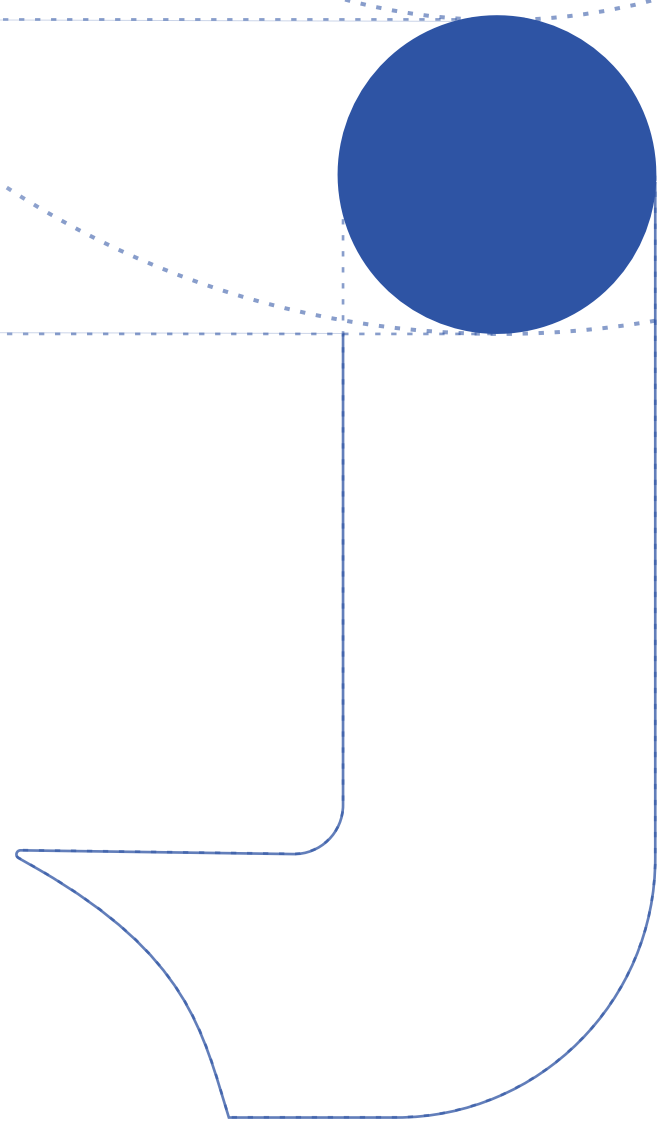
NLP

NATURAL
LANGUAGE
PROCESSING

Baidu 百度



ai.baidu.com

Natural Language Processing

Applications

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

Language Understanding

- Semantic Analysis
- Document Understanding
- Syntactic Analysis
- Information Extraction
- Lexical Analysis
- Sentiment Analysis

Language Generation

- Response Generation
- Comment Generation
- News Generation
- Question Generation
- Summarization
- 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

Application

- Web/Video Search
- News Feed
- Smart Speaker
- Document Intelligence
- Intelligent Customer Service

Platform

- EasyDL-NLP (Zero threshold AI Development Platform)
- BML-NLP (End-to-End AI Development Platform)
- ERNIE-Kit (End-to-end Development Toolkits)

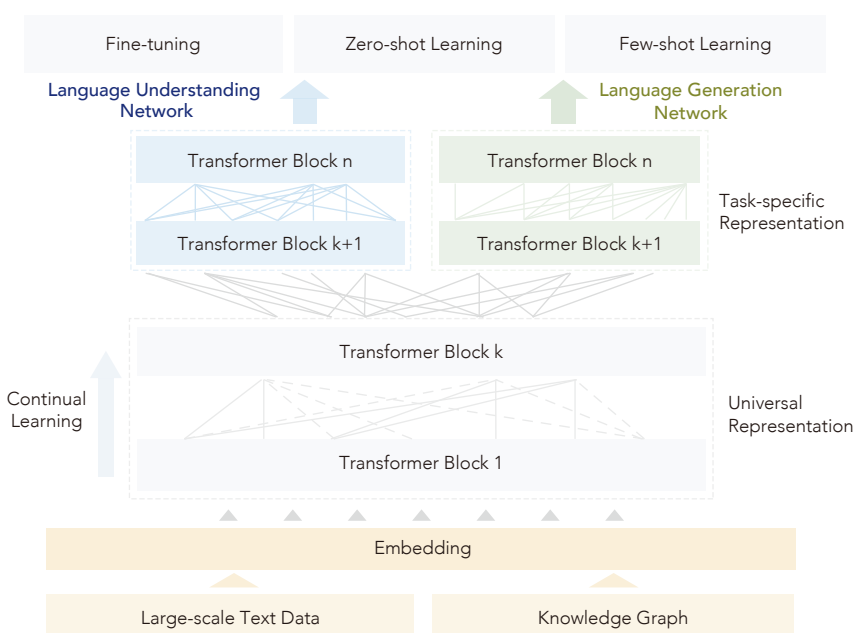
Model Family

- Cross-Modal
  - ERNIE-ViL (Vision-Language Understanding)
  - ERNIE-UNIMO (Unified -modality)
  - ERNIE-ViLG (Vision-Language Generation)
  - ERNIE-SAT (Speech -Language)
- Graph Learning
  - ERNIE-Sage (Text Graph Computing)
- Domain-specific Model
  - ERNIE-Health (Biomedicine)
  - ERNIE-Law (Law)
  - ERNIE-Finance (Finance)
- Task-specific Model
  - ERNIE-Gen (Language Generation)
  - ERNIE-Sim (Text Similarity)
  - ERNIE-IE (Information Extraction)
  - ERNIE-Search (Document Retrieval)
- Language Understanding & Generation
  - ERNIE 3.0 (10 Billion Parameters)
  - PCL-BAIDU Wenxin (260 Billion Parameters)
  - ERNIE 3.0 Zeus (Task Knowledge Enhanced Model)
- Cross-lingual Model
  - ERNIE-M (Cross-lingual model)

Framework

- Knowledge-enhanced Continual Pre-training Framework

PCL-BAIDU WENXIN: (ERNIE 3.0 TITAN) Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation



- Large Scale 260 billion parame-ter 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

PGL PaddlePaddle Graph Learning Framework

Winner of KDD CUP 2021 / TextGraph 2020 Graph Learning Challenges

Application	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	
	Hybrid Model					
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					

https://github.com/PaddlePaddle/PGL

PARL PaddlePaddle Reinforcement Learning Framework

Winner of NeurIPS 2018/2019/2020 Learning To Run Challenges

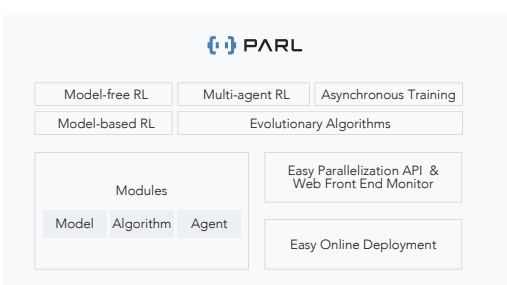
Reinforcement Learning Framework for Research & Applications

High quality reproduction of over 10+ mainstream RL methods

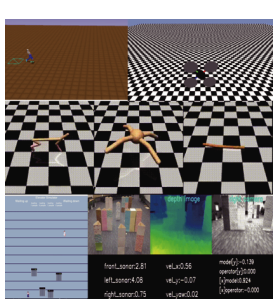
Easy implementation in CPU, GPU and clusters

Flexible framework, facilitating easy and low-cost secondary development

Applications



Research



https://github.com/paddlepaddle/PARL

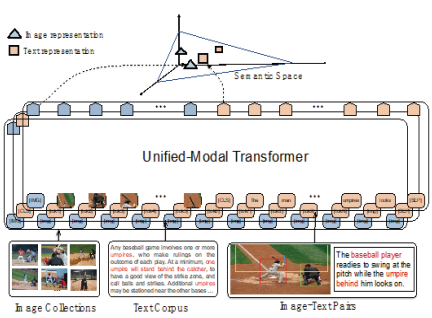
DOCUMENT UNDERSTANDING AND GENERATION

Applications	Recommendation	Search	Healthcare	Finance	Media	...
Tasks	Understanding: Topic, Sentiment, Opinion, Entity, Event, Relation			Generation: Video2Text, Text2Video, Text2Image, Summarization, Writing, Style Transfer		
Methods	Unified-Modal Pretraining			Unified Fine-tuning		

Unified Framework

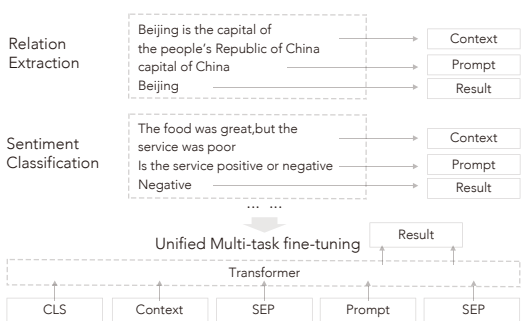
Unified-Modal Pre-training

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

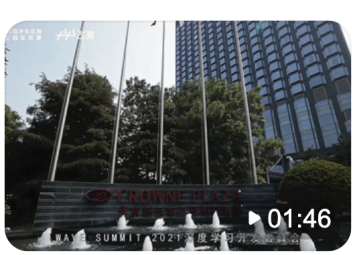
Support various NLU tasks through language prompts, and achieve good few-shot and zero-shot performances.



Application

Cross-Modal Tagging

Wave Summit 2021 for Deep Learning Developer



Entity : Baidu, PaddlePaddle
Event : Wave Summit 2021
Topic : Deep Learning, AI

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℃
    - 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

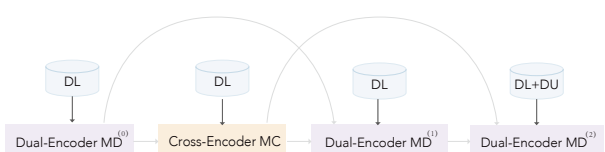
DUREADER Chinese Question Answering Datasets from Baidu Search

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] [DuReaderrobust: A Chinese Dataset Towards Evaluating Robustness and Generalization of Machine Reading Comprehension in Real-World Applications, In ACL 2021]
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 [DuReadervis: A Chinese Dataset for Open-domain Document Visual Question Answering, In ACL 2022]
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]

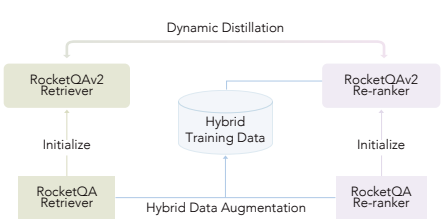
https://github.com/baidu/DuReader

ROCKETQA The Dense Retrieval Models for Open-domain Question Answering

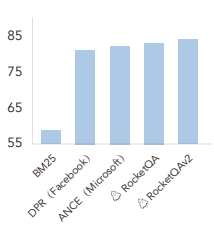
RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain QA, In NAACL2021



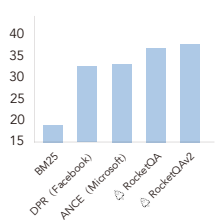
RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking, In EMNLP2021



Natural Questions



MSMARCO



All codes, models and tools are available at github.com/paddlepaddle/rocketqa

· State-of-the-art

· First-Chinese-model

· Easy-to-use

BAIDU TRANSLATE Communicate with the World

Products & Services

Document Translation | Webpage Translation | Video Translation

APP | API | SDK | PC-web | PC-client

Models & Algorithms

Multi-language	Multi-domain	Cross-modal
Multi-task Learning	Pretraining	Simultaneous Translation
Multi-agent Learning	Fine-tuning	ASR Error Correction
Data Augmentation	Adaptation	Interactive Decoding
Language Clustering	Ensemble	OCR Translation

Joint Training | Adversarial Training | Knowledge Distillation | Fast Decoding

Products

200+ Languages

500K+ Developers

100B+ Characters/Day

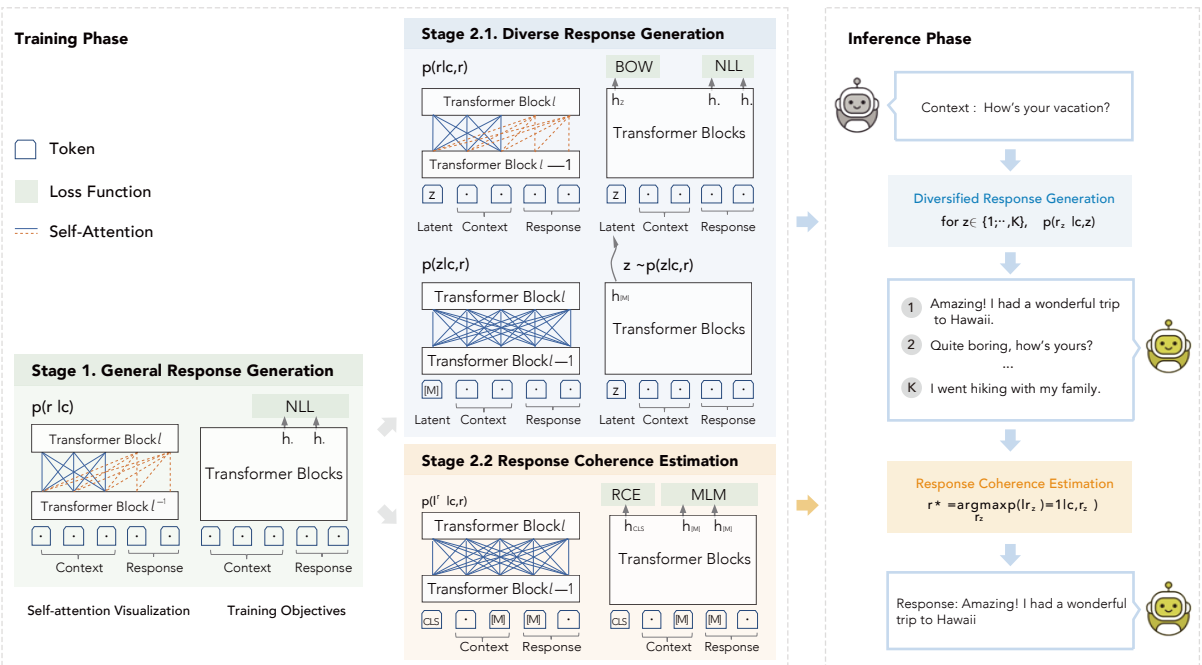
PC-client

Simultaneous Translation

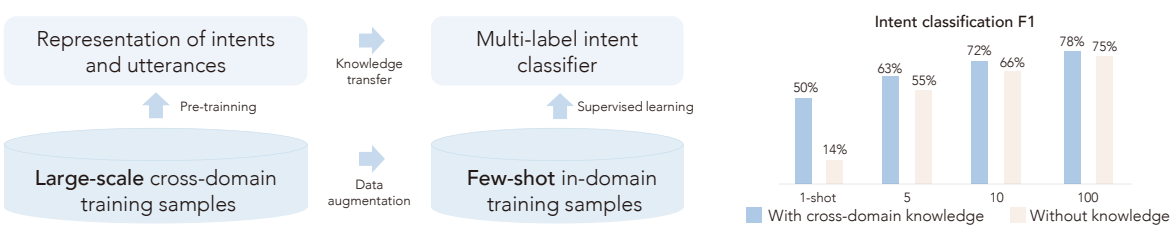
Document Translation

CONVERSATIONAL AI

PLATO: Large Scale Diversified Open-Domain Dialogue Generation



Task-Oriented Intent Classification with Cross-Domain Knowledge



Ecosystem of Technologies, Products and Services

Services	Proposal & Consulting	Training & Certification	Pre-delivery QC	Commissioned Development
Products	Intelligent Dialog Solutions			
	Customer Service	Consumer Electronics	Smart Home	Smart Travel
	UNIT: Customization and Service Platform for Conversational AI			Speech Technology Platform
	Skill and Knowledge Management			Domain Speech Solution / Self-service Training Platform
	Custom Skills :Task-dialog Q&A	Pre-built Skills	Chatbot	Knowledge Base / Wake Words / ASR / TTS
Technologies	Natural Language Processing	Knowledge Graph	Speech	

CLOUD API & OPEN-SOURCE LIBRARY

Cloud API Programmatic Interfaces to Baidu AI Platform services of NLP

Basic services	Application services
Lexical Analysis	Document Classification and Tagging
Dependency Parsing	Summarization
Semantic Similarity	Text Correction
Language Model	Sentiment Analysis
Word Embedding	Emotion Recognition

https://ai.baidu.com

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

Pre-built NLP Tasks

Information Extraction	Sentiment Analysis	Machine Translation	Machine Reading Comprehension	Knowledge Driven Dialogue	Text to SQL
Lexical Analysis	Syntactic Analysis	Language Model	Text Classification	Text Matching	Text Generation

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

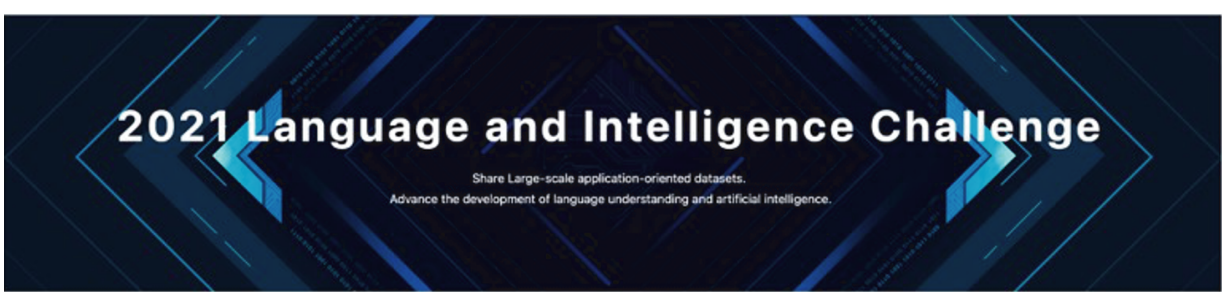
Initiated by Baidu, CCF and CIPSC, with the purpose of accelerating the advancement of NLP techniques, LUGE evaluates NLP models in terms of robustness and adaptability across multiple tasks and multiple domains. Now it has collected 11 tasks, consisting of 37 Chinese datasets by authors from 15 organizations.

NLP Tasks

Sentiment Analysis	Machine Reading Comprehension	Information Extraction	Dialogue
Text Similarity	Semantic Parser	Simultaneous Translation	Entity Linking

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 form academia

51% teams from industry and individuals

10,995 submissions

LIC · 2022 is on going

http://lic2022.cipsc.org.cn

NLP Openings
Full-time/Internship

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

BAIDU
NATURAL LANGUAGE PROCESSING

Email:nlp@baidu.com
Web:ai.baidu.com