利用文献资源库提升图书馆智慧服务能力

——科技文献知识人工智能引擎（SciAIEngine）的研究与实践

张智雄

中国科学院文献情报中心

2020年12月9日

提纲

- 知识获取能力：AI飞速突破的本质
- 科技文献库：图书馆智慧服务的一把钥匙
- SciAIEngine：智慧服务能力提升的思路
- SciAIEngine：智慧服务能力提升的实践
- 下一步的工作

提纲

- 知识获取能力：AI飞速突破的本质
- 科技文献库：图书馆智慧服务的一把钥匙
- SciAIEngine：智慧服务能力提升的思路
- SciAIEngine：智慧服务能力提升的实践
- 下一步的工作

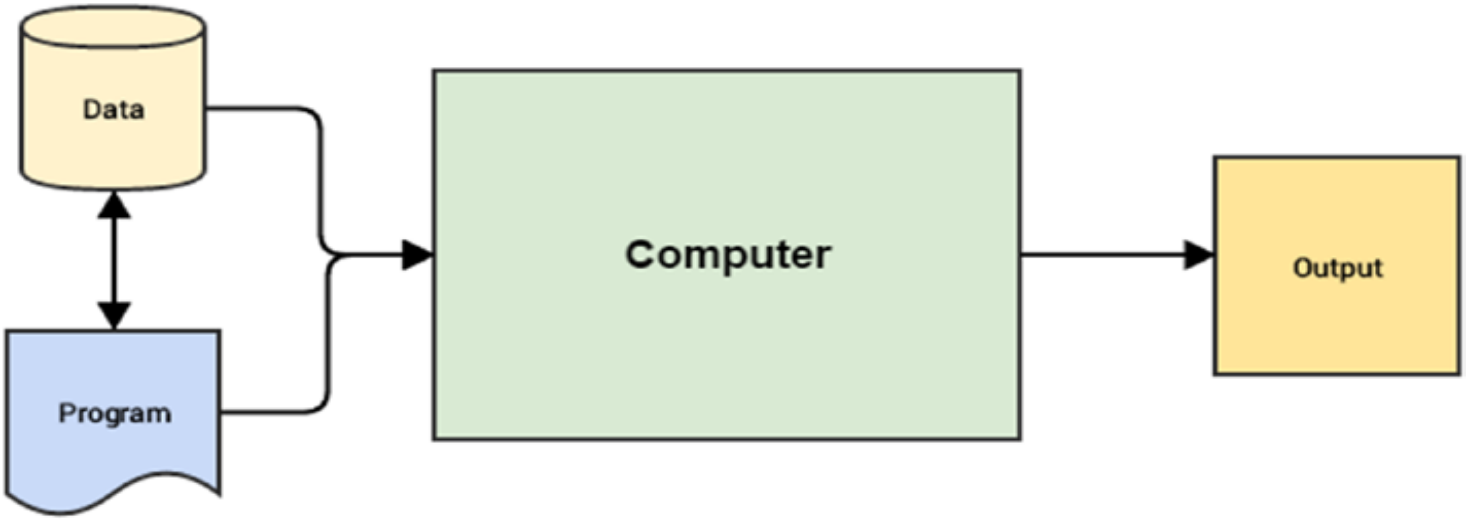
人工智能近来取得飞速突破

- 问答系统
  - 从文档中找到既定问题答案的准确率从2015年的60%提升至2017年的近80%，已经越来越接近人类
- 语音识别
  - 准确率在2017年已经提升至95%，达到人类水平。
  - 科大讯飞，语音识别
- 图像识别
  - 物体图像识别的错误率从2010年的28.5%下降到了2017年的2.5%，已超越人类水平。
  - “刷脸支付”（Paying With Your Face）

计算机解决问题的模式在改变

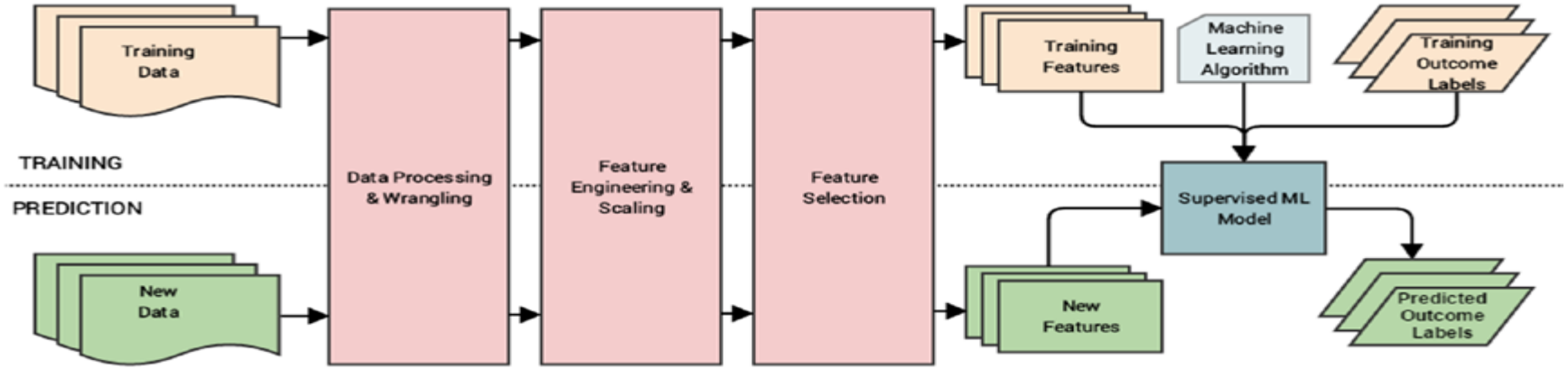
- 改变之一：
  - 从人输入知识让机器完成任务，到让机器学习知识，再让机器去完成任务

传统计算机程序解决问题模式：人输入知识让机器去完成任务



机器学习（Machine Learning）解决问题模式：让机器学习知识，再让机器去完成任务

- 模型训练阶段（Training）：利用标注好的数据语料，来训练模型，形成解决问题的知识
- 模型预测阶段（Predication）：解决问题阶段，利用训练好的模型（解决问题的知识），来解决类似的问题。



计算机解决问题的模式在改变

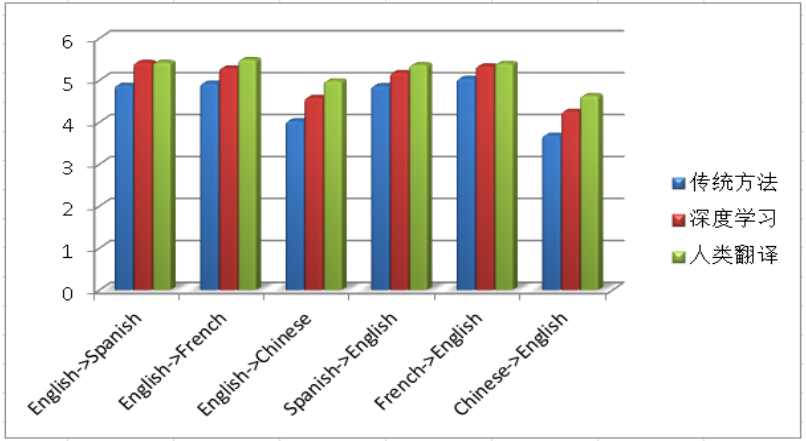
- 改变之一：
  - 从人输入知识让机器完成任务，到让机器学习知识，再让机器去完成任务
- 改变之二：
  - 拥有大样本训练语料和大规模计算能力，使得基于人工神经网络（artificial neural network）的深度学习（deep learning）的知识学习性能大幅提升

基于大样本训练语料和大规模计算能力人，使得深度学习的性能大幅提升

- **Google's Neural Machine Translation (GNMT)**
- 机器翻译。使用了深度学习的机器翻译算法在2016年把机器翻译的质量提高了58%-87%，迅速接近人类翻译水平
- **https://arxiv.org/pdf/1609.08144.pdf**

传统方法　　深度学习方法

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%



基于大样本训练语料和大规模计算能力人，使得深度学习的性能大幅提升

循环神经网络

Recurrent Neural Network, RNN

Long Short-Term memory (LSTM)

https://arxiv.org/pdf/1609.08144.pdf

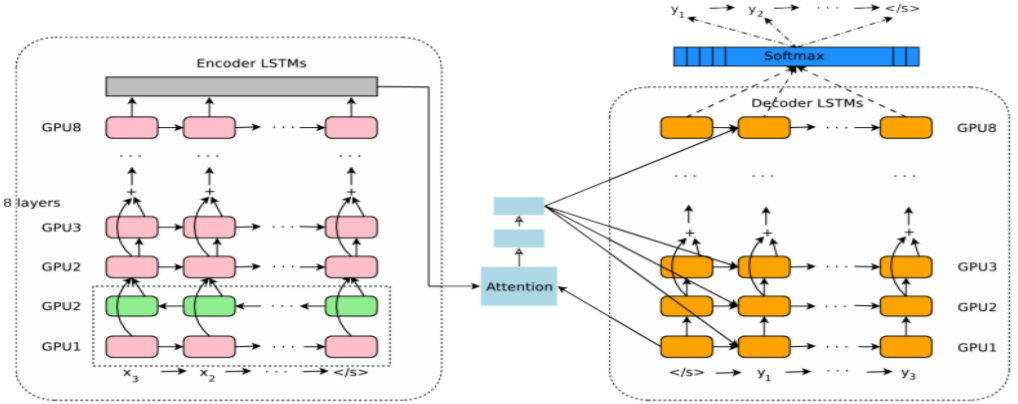


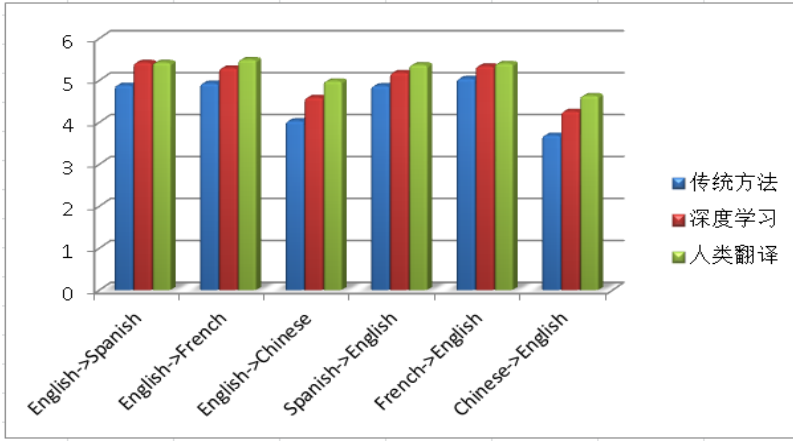
Figure 1: The model architecture of GNMT, Google's Neural Machine Translation system. On the left is the encoder network, on the right is the decoder network, in the middle is the attention module. The bottom encoder layer is bi-directional: the pink nodes gather information from left to right while the green nodes gather information from right to left. The other layers of the encoder are uni-directional. Residual connections start from the layer third from the bottom in the encoder and decoder. The model is partitioned into multiple GPUs to speed up training. In our setup, we have 8 encoder LSTM layers (1 bi-directional layer and 7 uni-directional layers), and 8 decoder layers. With this setting, one model replica is partitioned 8-ways and is placed on 8 different GPUs typically belonging to one host machine. During training, the bottom bi-directional encoder layers compute in parallel first. Once both finish, the uni-directional encoder layers can start computing, each on a separate GPU. To retain as much parallelism as possible during running the decoder layers, we use the bottom decoder layer output only for obtaining recurrent attention context, which is sent directly to all the remaining decoder layers. The softmax layer is also partitioned and placed on multiple GPUs. Depending on the output vocabulary size we either have them run on the same GPUs as the encoder and decoder networks, or have them run on a separate set of dedicated GPUs.

深度学习飞速突破的本质所在

- Google's Neural Machine Translation (GNMT)
- https://arxiv.org/pdf/1609.08144.pdf
- 数据集：隐藏着的是英文和法文之间的翻译知识
  - WMT dataset: Google internal production datasets
  - On WMT En-Fr, the training set contains 36M sentence pairs. （36,000,000个句子对），
  - On WMT En-De, the training set contains 5M sentence pairs. （5,000,000个句子对）

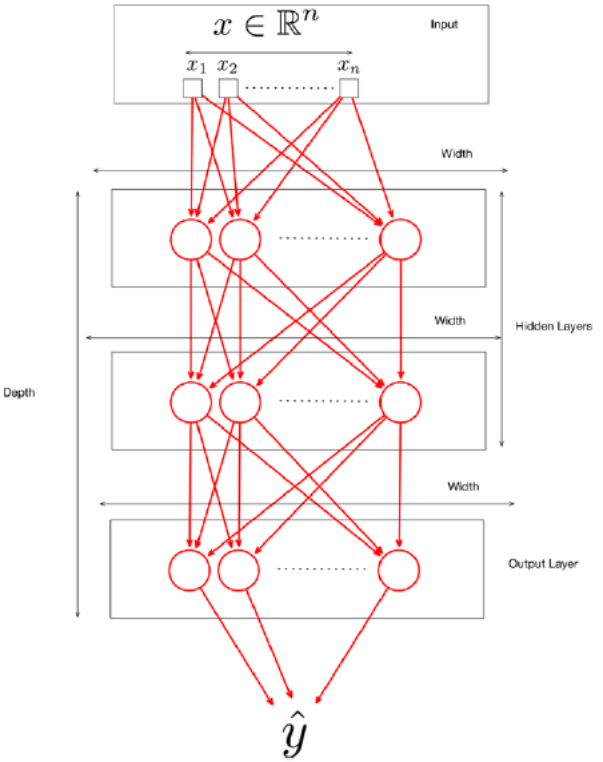
传统方法　　深度学习方法

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%



深度学习是一种自动学习各类语料特征的人工神经网络架构

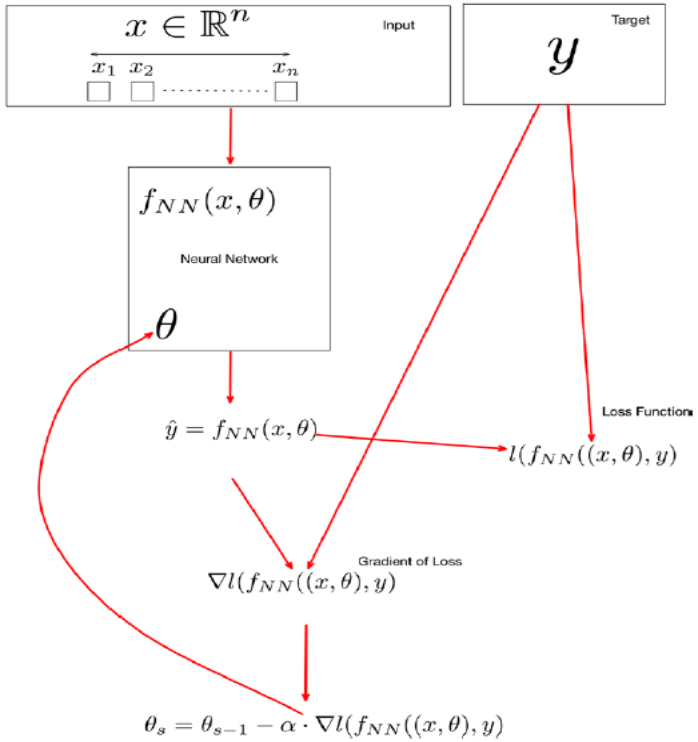
- The term **deep** in deep learning refers to the depth of the artificial neural network architecture
  - 深度一词是指人工神经网络架构的深度
- **Learning** stands for learning through the artificial neural network itself
  - 学习代表着通过人工神经网络本身进行特征学习



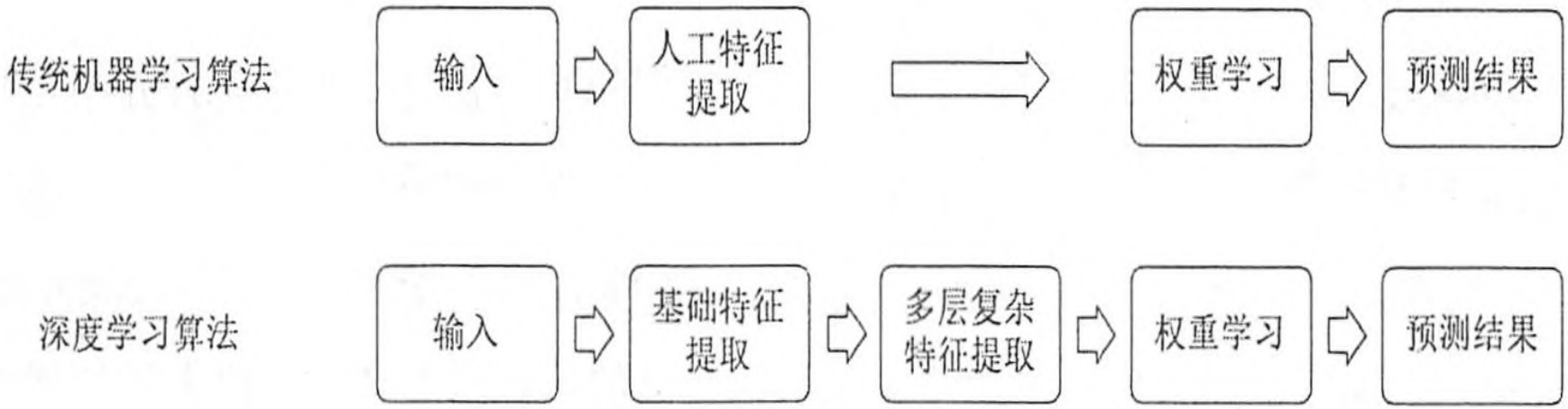
神经网络训练：“自动”形成一个基于能够对输入的数据进行目标结果预测的函数表示（Function Representation）

- θ 包括各类权重和偏移项的神经网络
- f_{NN} 整个神经网络的函数表示（模型）
- ^y，某个数据^x的神经网络输出
- l(^y, y) 损失函数，也即 (f_{NN}(x, θ), y)
- 损失函数的梯度 $\nabla l(f_{NN}(x,\theta),y)$
- 利用最新一轮的结果进一步优化θ

$$\theta_s = \theta_{s-1} - \alpha \cdot l(f_{NN}(x,\theta),y)$$



Machine learning/deep learning



Epoch
000,000

Learning rate
0.03

Activation
Tanh

Regularization
L1

Regularization rate
0.003

Problem type
Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

REGENERATE

FEATURES

Which properties do you want to feed in?

X_1
X_2
X_1^2
X_2^2
X_1X_2
$\sin(X_1)$
$\sin(X_2)$

2 HIDDEN LAYERS

4 neurons

2 neurons

*This is the output from one **neuron**. Hover to see it larger.*

*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

OUTPUT

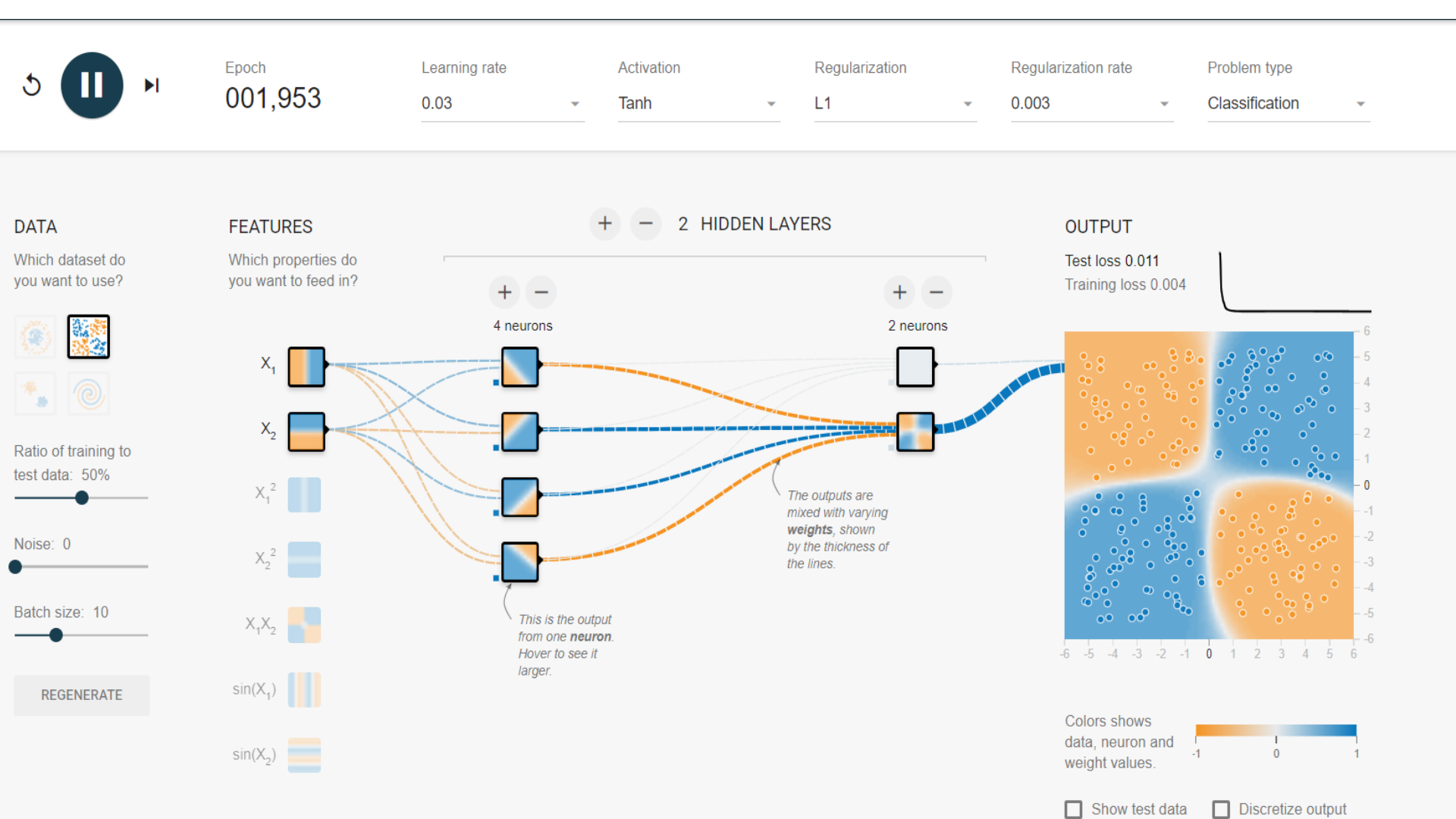
Test loss 0.509
Training loss 0.529

Colors shows data, neuron and weight values.

-1 0 1

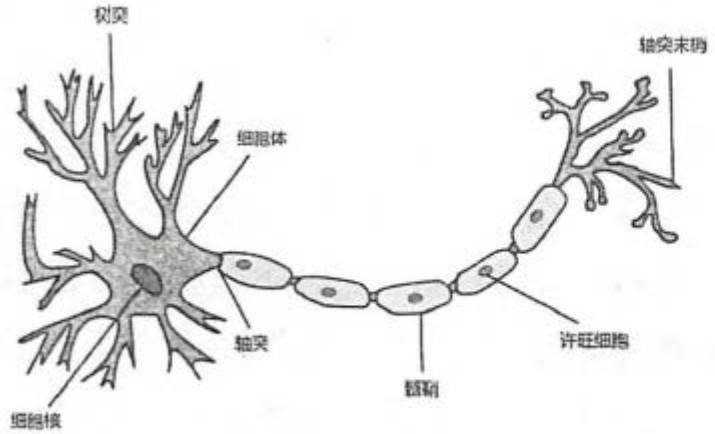
Show test data

Discretize output

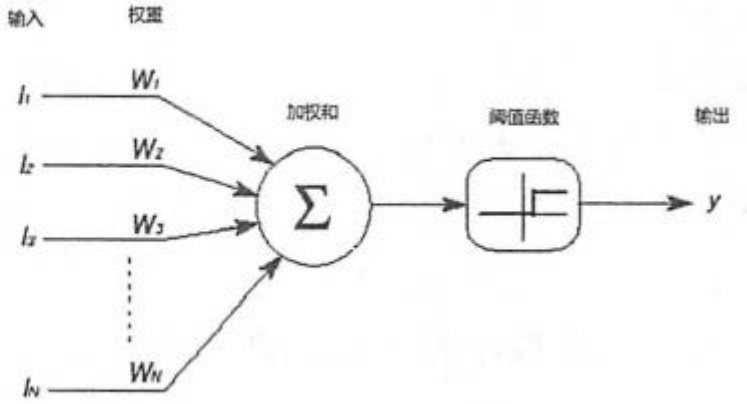

Epoch
001,953
Learning rate
0.03
Activation
Tanh
Regularization
L1
Regularization rate
0.003
Problem type
Classification
DATA
Which dataset do you want to use?
Ratio of training to test data: 50%
Noise: 0
Batch size: 10
REGENERATE
FEATURES
Which properties do you want to feed in?
X1
X2
X1²
X2²
X1X2
sin(X1)
sin(X2)
2 HIDDEN LAYERS
4 neurons
2 neurons
This is the output from one neuron. Hover to see it larger.
The outputs are mixed with varying weights, shown by the thickness of the lines.
OUTPUT
Test loss 0.011
Training loss 0.004
Colors shows data, neuron and weight values.
-1
0
1
Show test data
Discretize output

深度学习方法，近来取得重大应用突破。但其实它早就存在，但为什么一直沉睡直到今天?

- 1943年，最早的神经网络模型



（a）人类神经元结构

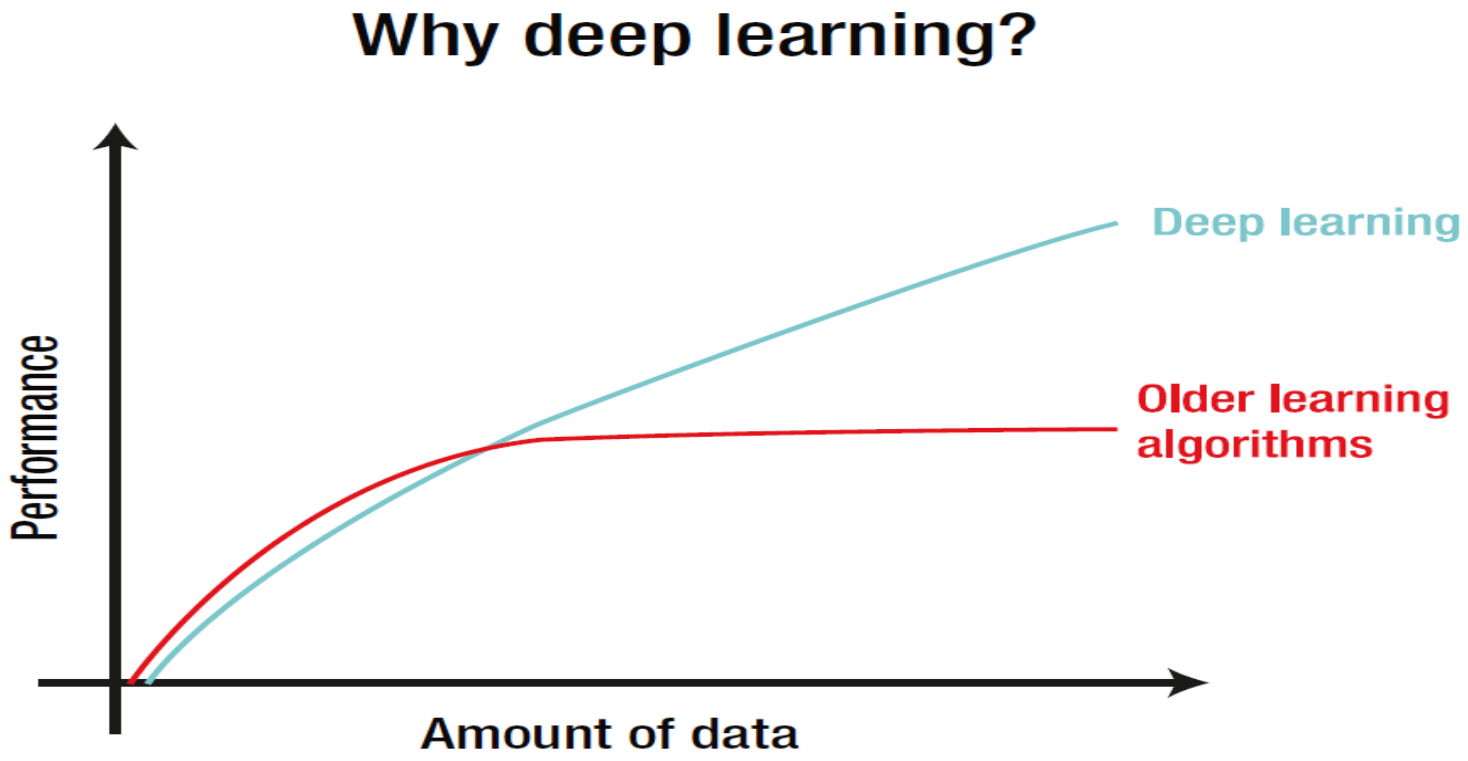


（b）McCulloch-Pitts Neuron 结构

深度学习，人工神经网络，早就提出

- 20世纪，在80年代末到90年代初，神经网络又得到高速发展，但最终没有盛行，败于传统机器学习
  - Sepp Hochreiter; Jürgen Schmidhuber 1991就已经提出，Long short-term memory (LSTM)
  - 常用的是依然是传统机器学习方法，如支持向量机（Support Vector Machine, SVM）等
  - 1998年，基于SVM的手写字母识别，错误率仅为0.8%，基于人工神经网络的深度学习达不到这样的精度

深度学习飞速突破的本质所在



深度学习飞速突破的本质总结

- 各类深度学习模型是基础
  - 人工神经网络（DNN、RNN、CNN）、Attention机制、Tansformer、Decoder、Encoder结构
- 大量可计算数据资源（训练语料）是前提
  - 隐藏着重要人类知识的大样本训练语料。语料是表征，实质是人类知识
- 大规模计算能力是催化剂
  - 高性能计算
    - CPU、GPU、Google's Tensor Processing Unit (TPU)

计算机解决问题的模式在改变

- 改变之一：
  - 从人输入知识让机器完成任务，到让机器学习知识，再让机器去完成任务
- 改变之二：
  - 拥有大样本训练语料和大规模计算能力，使得基于人工神经网络的深度学习的知识学习性能大幅提升
- 改变之三：
  - 基于预训练（Pre-Training）和微调（Fine-Tuning）的两阶段学习方法，重新写了自然语言处理（NLP）方式。预示着无监督的文本知识学习成为NLP的重要一环！

BERT模型

- Bidirectional Encoder Representations from Transformers (BERT)
  - 基于预训练语言模型的深度学习方法，重新写了自然语言处理的方式，两阶段方式处理NLP问题
  - 通过对大规模语料（Wikipedia, BookCorpus）的无监督预训练（Pre-training），得到通用的语言模型
  - 然后将语言模型通过微调（Fine-tuning）应用于NLP下游任务

基于预训练的语言模型

自 BERT 打破GLUE所有 11 项 NLP 记录后，基于预训练的语言模型获得大量关注并持续刷新GLUE 排行榜及问答系统SQuAD 排行榜

Rank	Name	Model
1	Facebook AI	RoBERTa
2	XLNet Team	XLNet-Large (ensemble)
3	Microsoft D365 AI & MSR AI	MT-DNN-ensemble
4	GLUE Human Baselines	GLUE Human Baselines
5	王玮	ALICE large ensemble (Alibaba DAMO NLF
6	Stanford Hazy Research	Snorkel MeTaL
7	XLM Systems	XLM (English only)
8	张倬胜	SemBERT
9	Danqi Chen	SpanBERT (single-task training)
10	Kevin Clark	BERT + BAM

General Language Understanding Evaluation benchmark (GLUE)

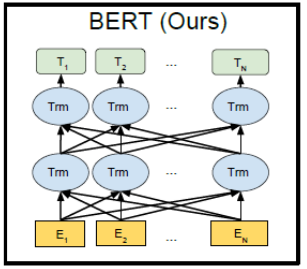
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Jul 19, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk*	88.050	90.645
3 Jul 23, 2019	XLNet + SG-Net Verifier (single model) *Shanghai Jiao Tong University & CloudWalk*	87.046	89.899
3 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
3 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
4 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147

Stanford Question Answering Dataset (SQuAD)

基于预训练的语言模型

- 基于 预训练+微调 模式的NLP预训练模型已成为主流
  - 预训练（Pre-Training）- 利用大规模无监督语料，学习语言特征（如：词法特征、句法特征、语法特征、上下文特征等）
  - 微调（Fine-tuning）- 针对具体下游任务（如：文本分类、命名实体识别、问答系统、阅读理解等），加入相关标注语料，调优

BERT预训练语言模型



- 利用Masked LM，掩盖部分词，学习词级别的上下文双向语义信息
- 利用Next Sentence Prediction，以相邻的两个句子为单位，学习句子级别的语义信息

- Rather than *always* replacing the chosen words with `[MASK]`, the data generator will do the following:
- 80% of the time: Replace the word with the `[MASK]` token, e.g., `my dog is hairy → my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy → my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy → my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

Input = `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]`

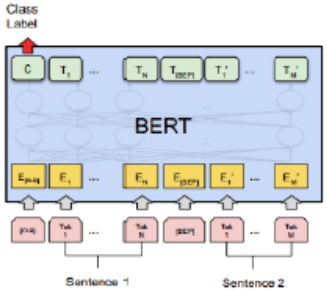
Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]`

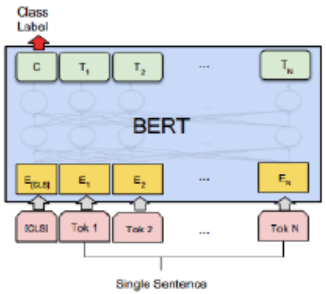
Label = `NotNext`

BERT 微调过程

句子关系类任务



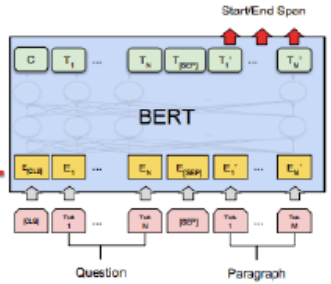
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



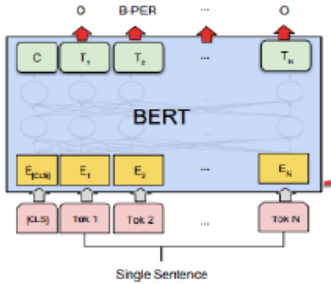
(b) Single Sentence Classification Tasks: SST-2, CoLA

单句分类任务

阅读理解任务



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

序列标注类任务

基于预训练的语言模型

语言模型	预训练语料	模型特点
ELMo (AllenNLP, 2018.02)	Word Benchmark	双向结构，LSTM神经网络
GPT (OpenAI, 2018.06)	BooksCorpus	单向结构，Transformers网络
BERT (Google, 2018.10)	BooksCorpus & Wikipedia	双向结构，Mask LM & NSP
GPT-2 (OpenAI, 2019.02)	WebText(45M links,40GB)	更大数据，可用于文本生成
SciBERT (AllenNLP, 2019.03)	1.14M papers from Semantic Scholar	针对英文科技文献
XLNet (CMU&Google, 2019.06)	BooksCorpus & Wikipedia & iga5，ClueWeb, Common Crawl	自回归、自编码
RoBERTa (Facebook AI, 2019.07)	BookCorpus, CC-news, OpenWebText, Stories	更大数据，更大批次，训练更久

基于BERT的中文预训练的语言模型

语言模型	预训练语料	模型特点
ERNIE 1.0 (百度, 2019.04)	中文维基百科（21M）、百度百科（51M）、百度新闻（47M）、百度贴吧（54M）	修改Mask策略，采用全词Mask、短语Mask、命名实体Mask 三种层级Mask策略
Chinese-BERT-wwm (哈工大&讯飞, 2019.06)	中文维基百科、新闻、问答	修改Mask策略，采用全词Mask
ERNIE 2.0 (百度, 2019.07)	百科、新闻、对话、搜索引擎数据	增加多种预训练，包括句子重排序、位置关系预测、语法关系、检索相关度等
OpenCLaP (清华大学, 2019.07)	民事文书（26M）、刑事文书（6M）	使用民事文书及刑事文书等领域语料
RoBERTa-zh (brightmart@github.com, 2019.09)	新闻、社区问答、多个百科数据	基于RoBERTa英文模型训练策略，利用大量中文语料开展训练

计算机解决问题的模式在改变

- 改变之一：
  - 从人输入知识让机器完成任务，到让机器学习知识，再让机器去完成任务
- 改变之二：
  - 拥有大样本训练语料和大规模计算能力，使得基于人工神经网络的深度学习的知识学习性能大幅提升
- 改变之三：
  - 基于预训练（Pre-Training）和微调（Fine-Tuning）的两阶段学习方法，重新写了自然语言处理（NLP）方式。预示着无监督的文本知识学习成为NLP的重要一环！

计算机解决问题的模式在改变

- 计算机解决问题三个模式的改变，都是围绕着计算机学习知识、开发利用知识的模式改变的。它明确告诉我们：

- 隐藏于各种数据资源（语料）中的知识获取能力提升是AI飞速突破的本质所在

提纲

- 知识获取能力：AI飞速突破的本质
- 科技文献库：图书馆智慧服务的一把钥匙
- SciAIEngine：智慧服务能力提升的思路
- SciAIEngine：智慧服务能力提升的实践
- 下一步的工作

人工智能取得飞速突破，文献情报机构也迫切需要提升图书馆智慧服务能力

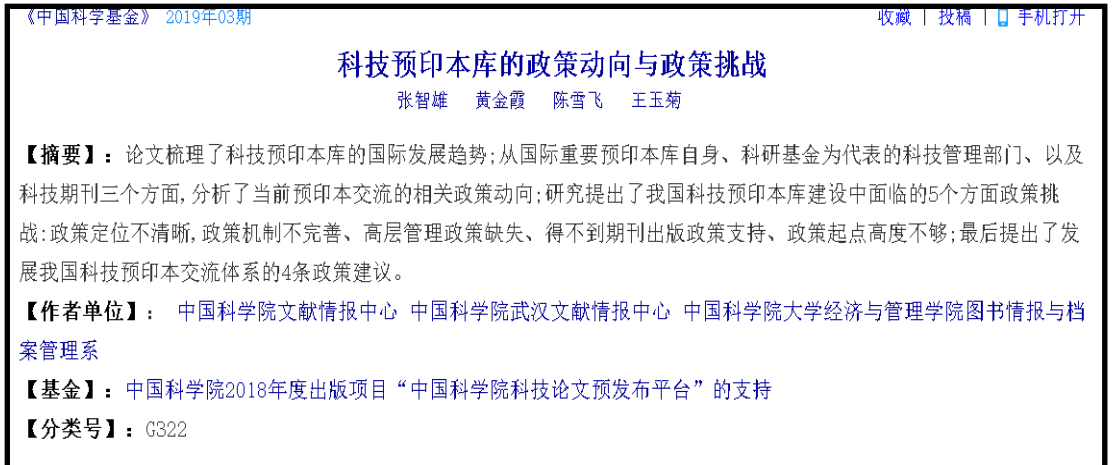
- 自动分类
  - 图书、文章、项目、专利、报告、人才、科技新闻分类…
- 自动文献内容标注
  - 段落、句子层面：研究问题、研究背景、主要方法、创新点
  - 短语、术语层面：关键词、理论、方法、工具…
- 自动推荐
  - 阅读推荐、评阅人推荐…
- 自动辅助阅读
  - 检索到2000篇相关论文…里面讲什么？

从科技文献（语料）中学习知识是提升图书馆智慧服务能力的钥匙

- 图书馆智慧服务需要各种上述AI能力
- 而上述AI需要语料来对模型进行训练，获取解决问题的知识
- 科技文献库（数字图书馆）就是最好的图书馆智慧服务的人工智能（AI）语料库！
  - 人类活动所形成的（文献作者、图书馆标引人员）
  - 隐藏丰富的文献知识语料

科技文献中的相关性知识

- 一篇科技论文，具有很多外部的特征，其中就隐藏着重要的知识关系，是很成熟的训练语料
  - 文献—作者，作者是某文本的标签....
  - 文献—期刊
  - 文献—机构
  - 文献—分类号
  - 文献—关键词



《中国科学基金》 2019年03期 收藏 | 投稿 | 手机打开

科技预印本库的政策动向与政策挑战

张智雄 黄金霞 陈雪飞 王玉菊

【摘要】：论文梳理了科技预印本库的国际发展趋势;从国际重要预印本库自身、科研基金为代表的科技管理部门、以及科技期刊三个方面,分析了当前预印本交流的相关政策动向;研究提出了我国科技预印本库建设中面临的5个方面政策挑战:政策定位不清晰,政策机制不完善、高层管理政策缺失、得不到期刊出版政策支持、政策起点高度不够;最后提出了发展我国科技预印本交流体系的4条政策建议。

【作者单位】： 中国科学院文献情报中心 中国科学院武汉文献情报中心 中国科学院大学经济与管理学院图书情报与档案管理系

【基金】：中国科学院2018年度出版项目“中国科学院科技论文预发布平台”的支持

【分类号】：G322

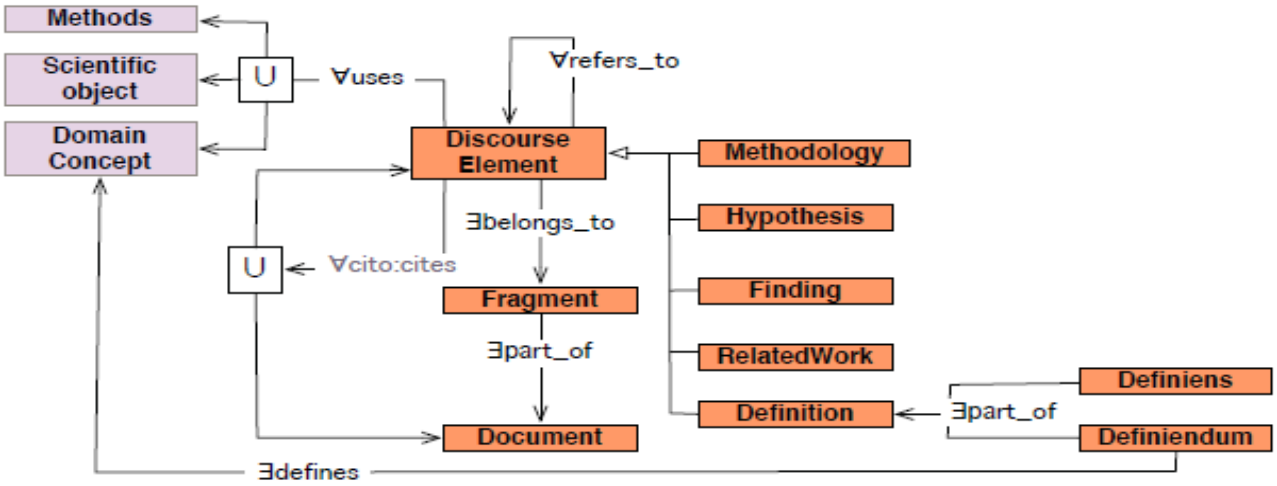
科技文献中的丰富语义知识

- 丰富语义（Rich Semantics）知识
  - 丰富语义（Rich Semantics）相对于一般意义上的语义（Semantics）而言，它是由多类型语义元素有机组合在一起的复合体，具有结构化、模型化的特征
  - 语义丰富化技术正在从文献中零散知识点及其关系的标注和揭示，向着更具有应用价值的、有框架和模型支撑的丰富语义抽取和揭示的方向发展

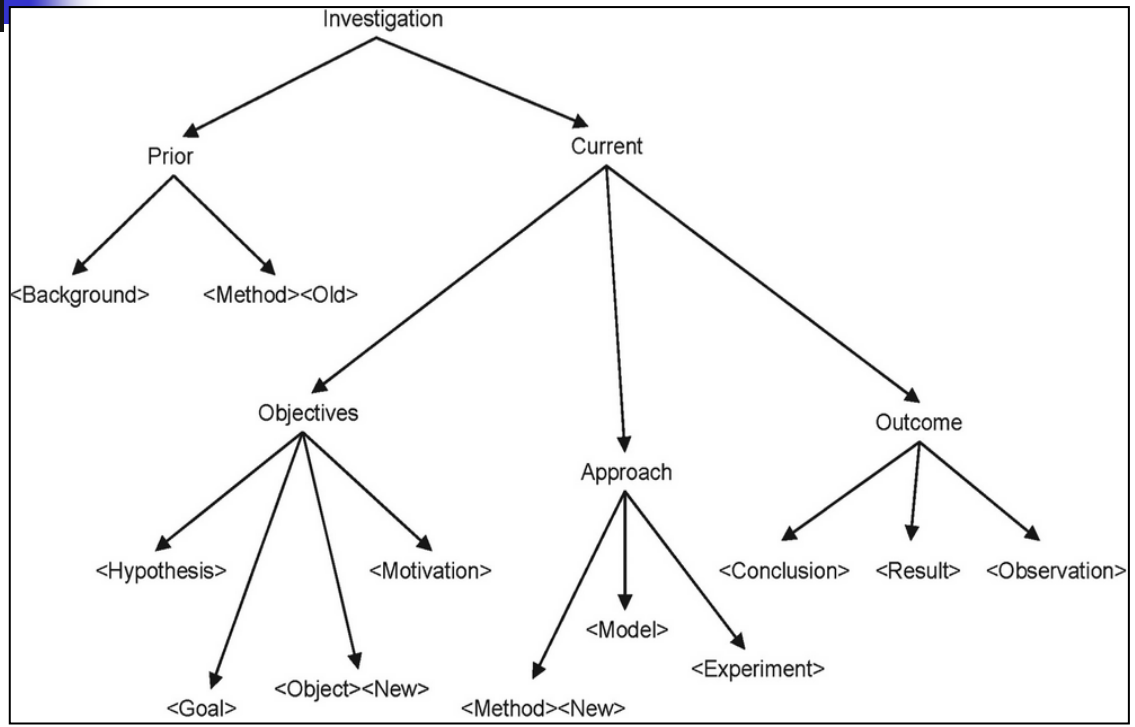
SciAnnotDoc模型揭示的丰富语义知识

H´el`ene de Ribaupierre and Gilles Falquet （2015），An Automated Annotation Process for the SciDocAnnot Scientific Document Model

Fig. 1. SciAnnotDoc model



CoreSC模型揭示的丰富语义知识



CoreSC（Core Scientific Concepts）

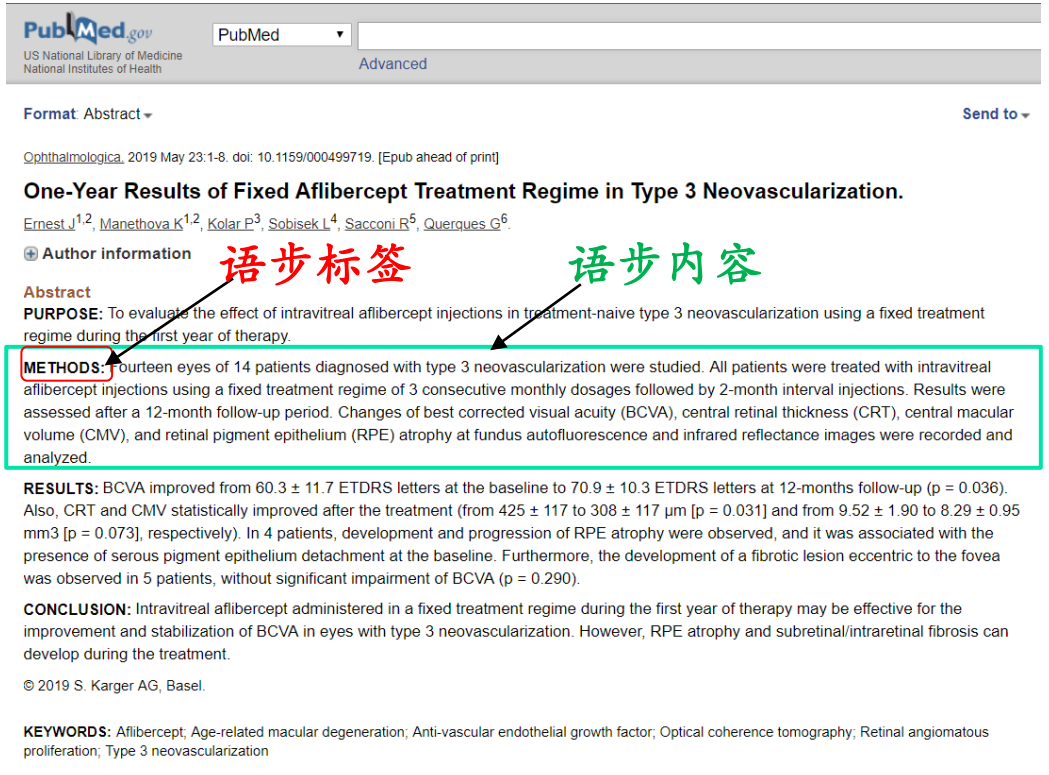
Liakata et al. (2010)

已标注好的语料随处可见

- 语步（Move）

科技论文摘要语步

科技论文的摘要中，各个语步作为相对独立的功能模块有机衔接起来，形成摘要的一个宏观的结构



PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed ▾ Advanced

Format: Abstract ▾ Send to ▾

Ophthalmologica. 2019 May 23:1-8. doi: 10.1159/000499719. [Epub ahead of print]

One-Year Results of Fixed Aflibercept Treatment Regime in Type 3 Neovascularization.

Ernest J[1,2], Manethova K[1,2], Kolar P[3], Sobisek L[4], Sacconi R[5], Querques G[6].

Author information

语步标签

语步内容

Abstract

PURPOSE: To evaluate the effect of intravitreal aflibercept injections in treatment-naive type 3 neovascularization using a fixed treatment regime during the first year of therapy.

METHODS: Fourteen eyes of 14 patients diagnosed with type 3 neovascularization were studied. All patients were treated with intravitreal aflibercept injections using a fixed treatment regime of 3 consecutive monthly dosages followed by 2-month interval injections. Results were assessed after a 12-month follow-up period. Changes of best corrected visual acuity (BCVA), central retinal thickness (CRT), central macular volume (CMV), and retinal pigment epithelium (RPE) atrophy at fundus autofluorescence and infrared reflectance images were recorded and analyzed.

RESULTS: BCVA improved from 60.3 ± 11.7 ETDRS letters at the baseline to 70.9 ± 10.3 ETDRS letters at 12-months follow-up (p = 0.036). Also, CRT and CMV statistically improved after the treatment (from 425 ± 117 to 308 ± 117 μm [p = 0.031] and from 9.52 ± 1.90 to 8.29 ± 0.95 mm3 [p = 0.073], respectively). In 4 patients, development and progression of RPE atrophy were observed, and it was associated with the presence of serous pigment epithelium detachment at the baseline. Furthermore, the development of a fibrotic lesion eccentric to the fovea was observed in 5 patients, without significant impairment of BCVA (p = 0.290).

CONCLUSION: Intravitreal aflibercept administered in a fixed treatment regime during the first year of therapy may be effective for the improvement and stabilization of BCVA in eyes with type 3 neovascularization. However, RPE atrophy and subretinal/intraretinal fibrosis can develop during the treatment.

KEYWORDS: Aflibercept; Age-related macular degeneration; Anti-vascular endothelial growth factor; Optical coherence tomography; Retinal angiomatous proliferation; Type 3 neovascularization

从科技文献（语料）中学习知识是图书馆智慧服务能力的钥匙

- 应当充分挖掘隐藏着丰富知识内容的科技文献资源
- 实现从“科技文献库”到“科技知识引擎”的转变

提纲

- 知识获取能力：AI飞速突破的本质
- 科技文献库：图书馆智慧服务的一把钥匙
- **SciAIEngine：智慧服务能力提升的思路**
- SciAIEngine：智慧服务能力提升的实践
- 下一步的工作

NSTL和中科院项目支持

- 国家科技图书文献中心(NSTL)专项课题
  - “下一代开放知识服务平台总体设计及关键技术研发”专项（2019）
  - “下一代开放知识服务平台关键技术优化集成与系统研发”专项（2020）
- 中国科学院文献情报能力建设专项课题
  - “基于科技文献知识的人工智能（AI）引擎建设”
  - “科技文献丰富语义检索应用示范”

- 致力于通过深度学习等技术方法，将隐藏于科技文献库中的语料资源转化成为解决问题的知识方案

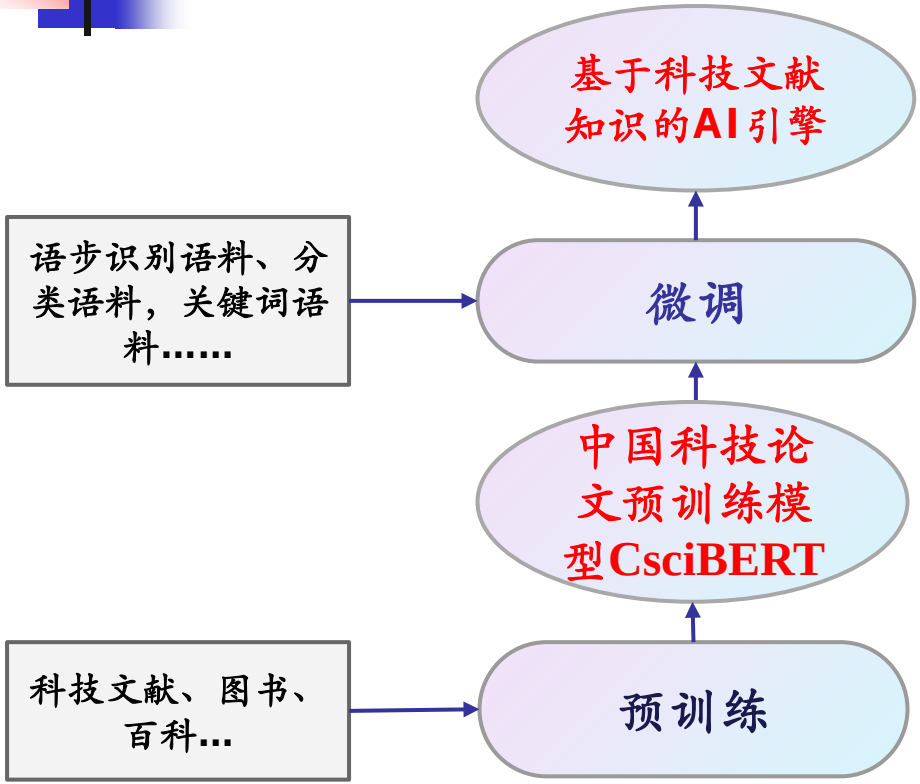
SciAIEngine：建设思路

- 充分和挖掘好隐藏着丰富知识内容的科技文献资源
- 将人类活动所形成的各类科技文献库看成一个隐藏丰富文献知识的语料
- 实现从“科技文献库”到“科技知识引擎”的转变

SciAIEngine：建设思路

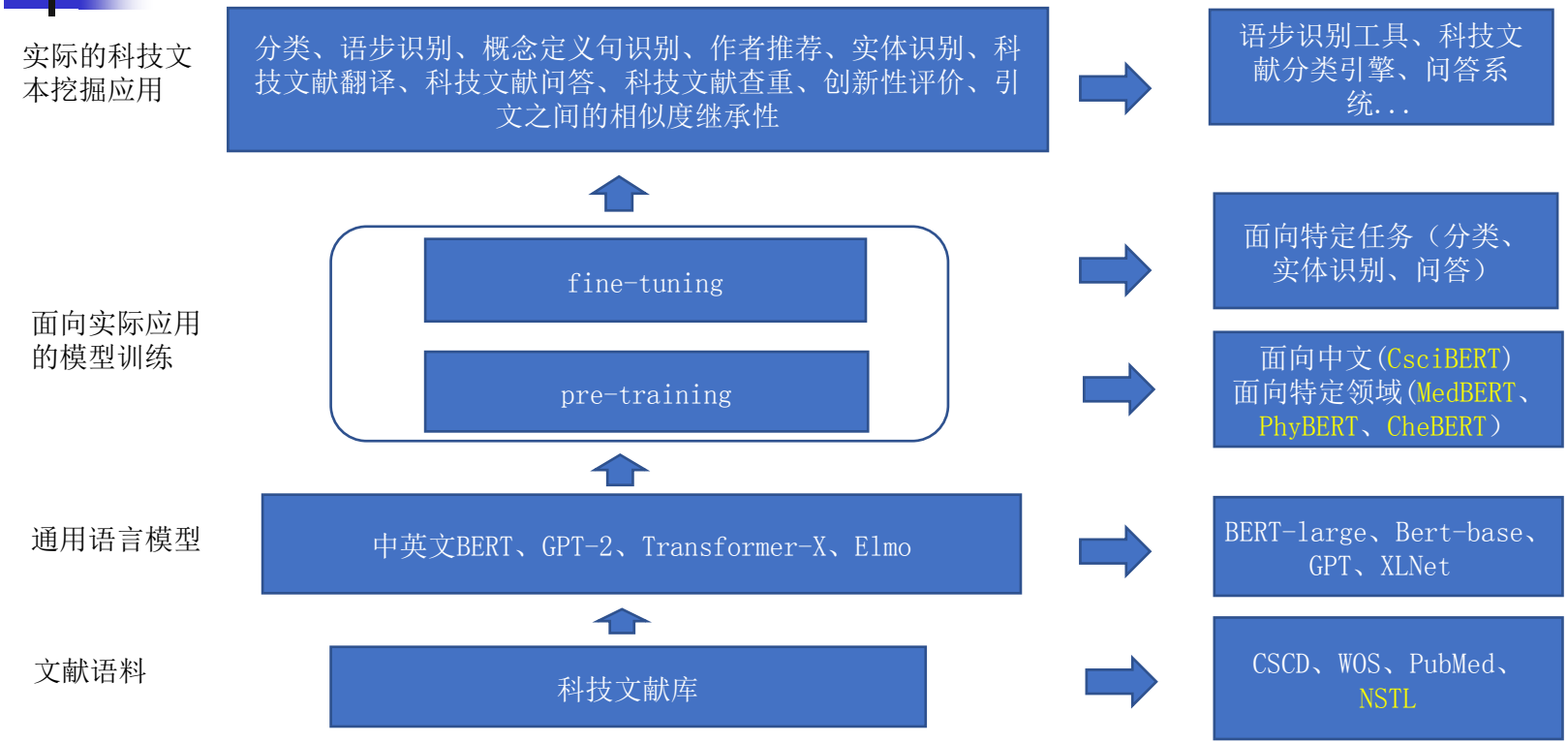
- 向全球人工智能研究和应用领域提供中文科技文献的基础语言模型和知识引擎
  - 建成并发布中国科技论文预训练模型(CsciBERT)
  - 建成并发布基于科技文献知识的AI引擎

从隐藏着丰富知识内容的科技文献中学习知识，并应用到需要解决问题的应用之中

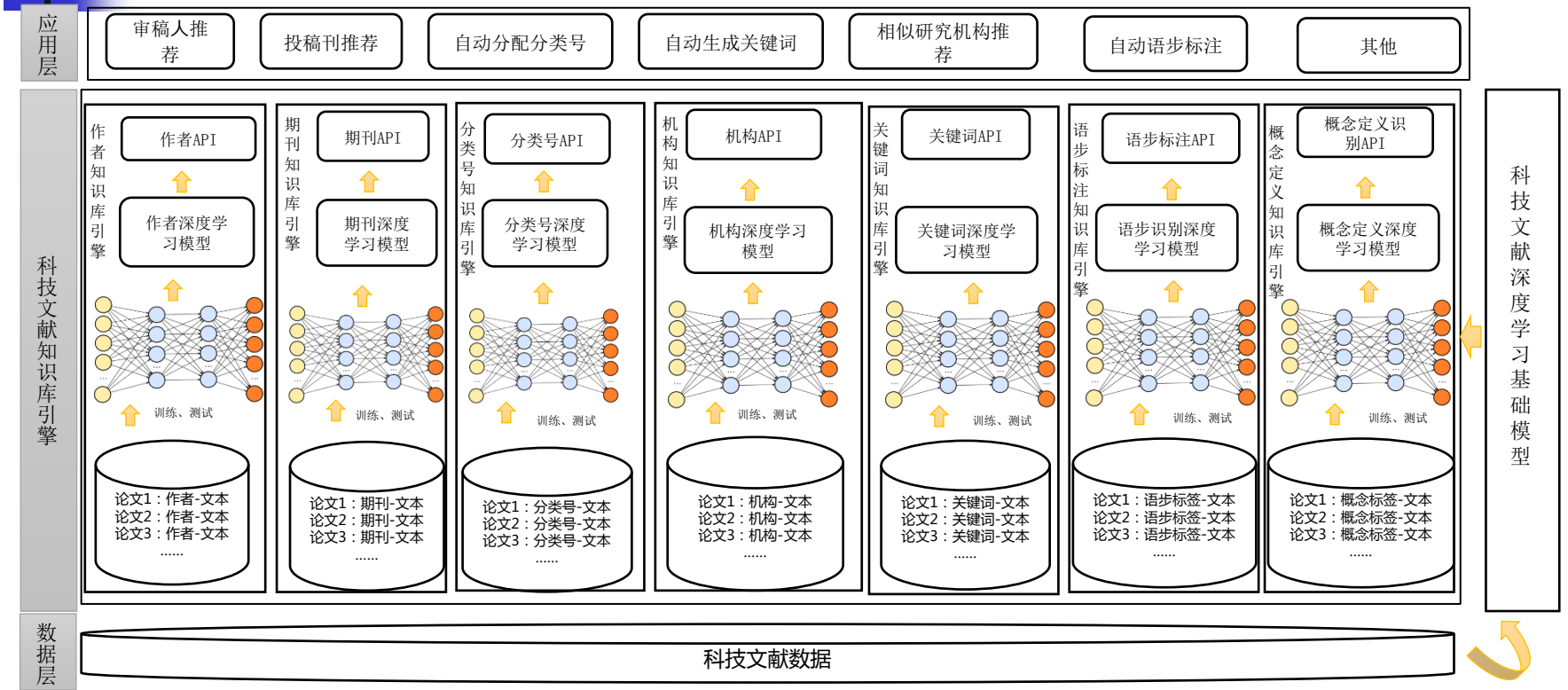


- 基于“预训练”（Pre-Training）和“微调”（Fine-tuning）的两阶段学习方法
- 从文献中学习知识，并提供引擎服务

中国科技论文预训练模型(CsciBERT)



基于科技文献知识的AI引擎



提纲

- 知识获取能力：AI飞速突破的本质
- 科技文献库：图书馆智慧服务的一把钥匙
- SciAIEngine：智慧服务能力提升的思路
- SciAIEngine：智慧服务能力提升的实践
- 下一步的工作

SciAIEngine：智慧服务能力提升的实践

- SciAIEngine是什么？
- SciAIEngine解决什么问题?
- SciAIEngine怎样用？
- SciAIEngine都在哪里用了？

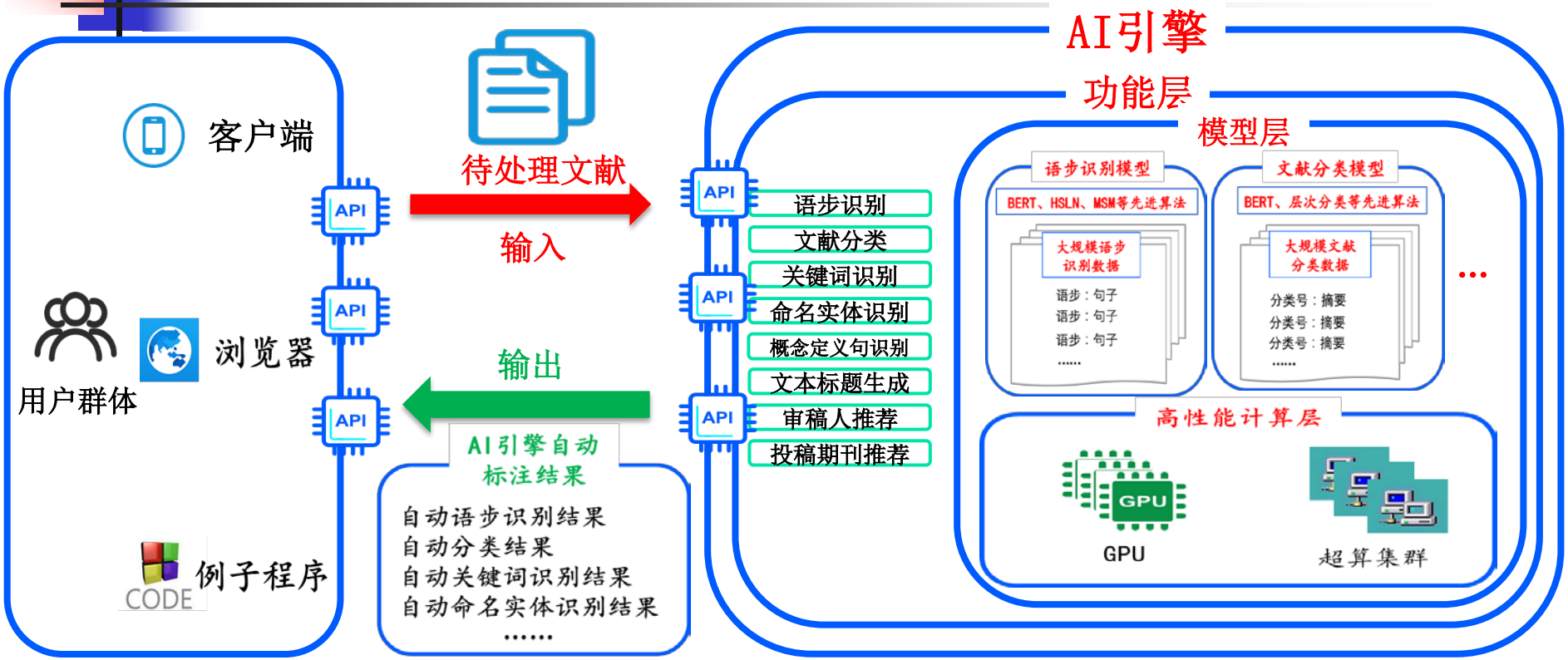
SciAIEngine：智慧服务能力提升的实践

- SciAIEngine是什么？
- SciAIEngine解决什么问题?
- SciAIEngine怎样用？
- SciAIEngine都在哪里用了？

科技文献知识AI引擎

- 科技文献知识AI引擎（SciAIEngine），是一款科技文献知识驱动的人工智能（AI）引擎。它利用科技文献大数据和深度学习技术方法，从科技文献中自动学习获取科技文本挖掘的重要知识，并基于这些知识构建起核心的人工智能组件，支撑科技文献的深入挖掘和利用

科技文献知识AI引擎的工作原理



SciAIEngine：智慧服务能力提升的实践

- SciAIEngine是什么？
- SciAIEngine解决什么问题？
- SciAIEngine怎样用？
- SciAIEngine都在哪里用了？

AI引擎解决什么问题？

- 提供迫切需要的
  - 科技文献挖掘
  - 计算机辅助阅读
  - 数据自动化处理
  - 情报智能化分析
  - ……
  - 人工智能核心组件

SciAIEngine功能

- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

语步识别

- 自动识别科技文献中的研究背景、研究目的、研究方法、研究结果、研究结论等重要句子，显性地揭示科技文献的重要内容

nature

Explore our content ⌄ Journal information ⌄

nature > articles > article

Article | Published: 02 December 2020

Autonomous navigation of stratospheric balloons using reinforcement learning

Marc G. Bellemare ✉, Salvatore Candido ✉, Pablo Samuel Castro, Jun Gong, Marlos C. Machado, Subhodeep Moitra, Sameera S. Ponda & Ziyu Wang

Nature **588**, 77–82(2020) | Cite this article

192 Altmetric | Metrics

Abstract

Efficiently navigating a superpressure balloon in the stratosphere[1] requires the integration of a multitude of cues, such as wind speed and solar elevation, and the process is complicated by forecast errors and sparse wind measurements. Coupled with the need to make decisions in real time, these factors rule out the use of conventional control techniques[2,3]. Here we describe the use of reinforcement learning[4,5] to create a high-performing flight controller. Our algorithm uses data augmentation[6,7] and a self-correcting design to overcome the key technical challenge of reinforcement learning from imperfect data, which has proved to be a major obstacle to its application to physical systems[8]. We deployed our controller to station Loon superpressure balloons at multiple locations across the globe, including a 39-day controlled experiment over the Pacific Ocean. Analyses show that the controller outperforms Loon's previous algorithm and is robust to the natural

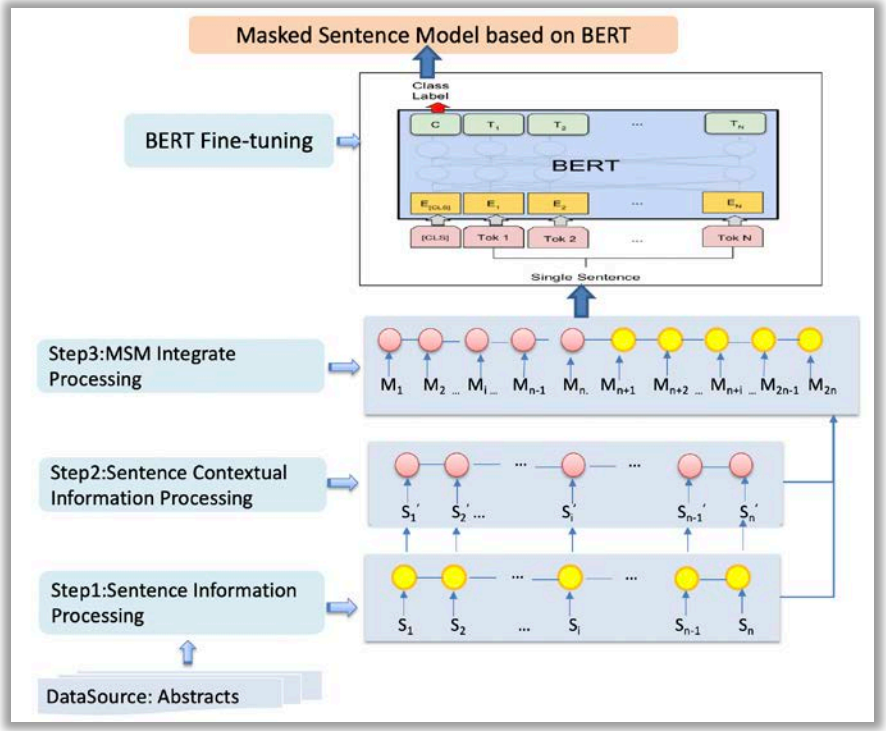
提出Masked Sentence Model等语步识别模型

- PubMed 20K RCT数据集评测结果

模型名称	提出者	模型效果F1值
BERT-Base Model	谷歌	86.05
HSLN Model	麻省理工学院	92.6
Masked Sentence Model	**课题组**	**91.15**
Masked Labels Model	**课题组**	**91.29**
Refined Masked Sentence Model	**课题组**	**93.21**

Masked Sentence Model

- 以句子本身学习其内容特征(content information)
- 在摘要中掩盖（Mask）某句子，以学习其上下文特征(context information)
- 融合两种模型输入，同时学习内容特征与上下文特征



SciAIEngine提供的语步识别引擎

	语步类型	模型方法	语料规模
英文摘要语步识别	Background、Objectives、Methods、Results、Conclusions	Refined Masked Sentence Model	**28万篇精炼**的英文结构化摘要
中文摘要语步识别	目的、方法、结果、结论	BERT fine-tuning	**80万篇**中文结构化摘要
基金项目语步识别	背景及问题、目标及任务、方法内容、价值意义	BERT fine-tuning	**2万条人工筛选**的句子语料

英文摘要语步识别引擎

- 英文非结构化摘要

Molecular and clinical analysis of 27 German patie with Leber congenital amaurosis

Nicole Weisschuh [1], Britta Feldhaus [1], Muhammad Imran Khan [2], Frans P M Cremers [2 3], Susanne Kohl [1], Bernd Wissinger [1], Ditta Zobor [1]

Affiliations + expand

PMID: 30576320 PMCID: PMC6303042 DOI: 10.1371/journal.pone.0205380

Free PMC article

Abstract

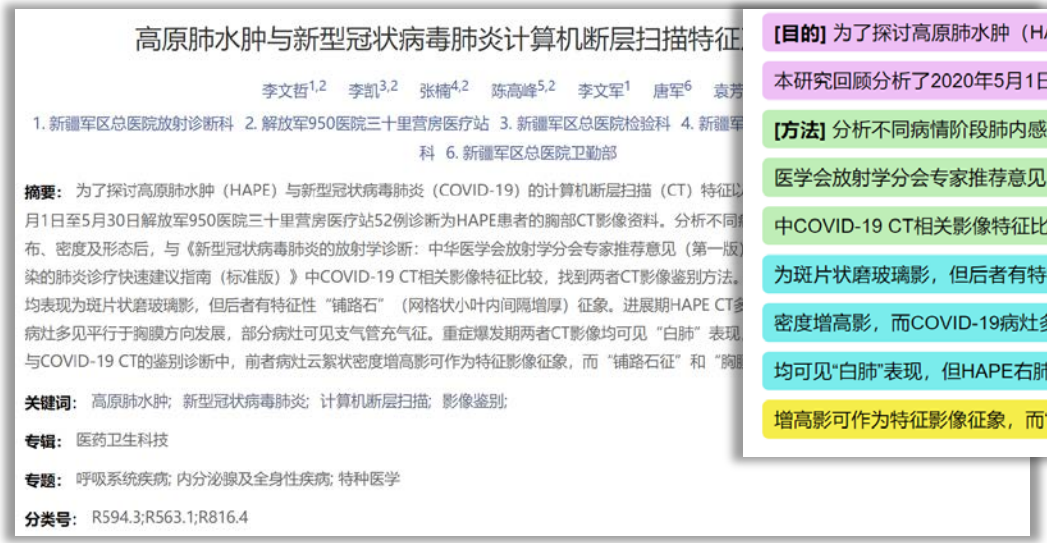
Leber congenital amaurosis (LCA) is the earliest and most severe form of all inherited retinal dystrophies (IRD) and the most frequent cause of inherited blindness in children. The phenotypic overlap with other early-onset and severe IRDs as well as difficulties associated with the ophthal examination of infants can complicate the clinical diagnosis. To date, 25 genes have been implica in the pathogenesis of LCA. The disorder is usually inherited in an autosomal recessive fashion, although rare dominant cases have been reported. We report the mutation spectra and frequen genes in 27 German index patients initially diagnosed with LCA. A total of 108 LCA- and other g implicated in IRD were analysed using a cost-effective targeted next-generation sequencing

- 语步识别自动标注结果

[BACKGROUND] Leber congenital amaurosis (LCA) is the earliest and most severe form of all inherited retinal dystrophies (IRD) and the most frequent cause of inherited blindness in children. **[BACKGROUND]** The phenotypic overlap with other early-onset and severe IRDs as well as difficulties associated with the ophthalmic examination of infants can complicate the clinical diagnosis. **[BACKGROUND]** To date, 25 genes have been implicated in the pathogenesis of LCA. **[BACKGROUND]** The disorder is usually inherited in an autosomal recessive fashion, although rare dominant cases have been reported. **[OBJECTIVES]** We report the mutation spectra and frequency of genes in 27 German index patients initially diagnosed with LCA. **[METHODS]** A total of 108 LCA- and other genes implicated in IRD were analysed using a cost-effective targeted next-generation sequencing procedure based on molecular inversion probes (MIPs). **[RESULTS]** Sequencing and variant filtering led to the identification of putative pathogenic variants in 25 cases, thereby leading to a detection rate of 93%. **[RESULTS]** The mutation spectrum comprises 34 different alleles, 17 of which are novel. **[CONCLUSIONS]** In line with previous studies, the genetic results led to a revision of the initial clinical diagnosis in a substantial proportion of cases, demonstrating the importance of genetic testing in IRD. **[CONCLUSIONS]** In addition, our detection rate of 93% shows that MIPs are a cost-efficient and sensitive tool for targeted next-generation sequencing in IRD.

中文摘要语步识别引擎

- 中文非结构化摘要
- 语步识别自动标注结果



高原肺水肿与新型冠状病毒肺炎计算机断层扫描特征

李文哲[1,2] 李凯[3,2] 张楠[4,2] 陈高峰[5,2] 李文军[1] 唐军[6] 袁芳

1. 新疆军区总医院放射诊断科 2. 解放军950医院三十里营房医疗站 3. 新疆军区总医院检验科 4. 新疆军科 6. 新疆军区总医院卫勤部

摘要： 为了探讨高原肺水肿（HAPE）与新型冠状病毒肺炎（COVID-19）的计算机断层扫描（CT）特征以月1日至5月30日解放军950医院三十里营房医疗站52例诊断为HAPE患者的胸部CT影像资料。分析不同布、密度及形态后，与《新型冠状病毒肺炎的放射学诊断：中华医学会放射学分会专家推荐意见（第一版）》和《新型冠状病毒（2019-nCoV）感染的肺炎诊疗快速建议指南（标准版）》中COVID-19 CT相关影像特征比较，找到两者CT影像鉴别方法。均表现为斑片状磨玻璃影，但后者有特征性“铺路石”（网格状小叶内间隔增厚）征象。进展期HAPE CT多病灶多见平行于胸膜方向发展，部分病灶可见支气管充气征。重症爆发期两者CT影像均可见“白肺”表现与COVID-19 CT的鉴别诊断中，前者病灶云絮状密度增高影可作为特征影像征象，而“铺路石征”和“胸

关键词： 高原肺水肿；新型冠状病毒肺炎；计算机断层扫描；影像鉴别；

专辑： 医药卫生科技

专题： 呼吸系统疾病；内分泌腺及全身性疾病；特种医学

分类号： R594.3;R563.1;R816.4

[目的] 为了探讨高原肺水肿（HAPE）与新型冠状病毒肺炎（COVID-19）的计算机断层扫描（CT）特征以及影像鉴别，

本研究回顾分析了2020年5月1日至5月30日解放军950医院三十里营房医疗站52例诊断为HAPE患者的胸部CT影像资料。

[方法] 分析不同病情阶段肺内感染病灶的数量、位置、分布、密度及形态后，与《新型冠状病毒肺炎的放射学诊断：中华医学会放射学分会专家推荐意见（第一版）》和《新型冠状病毒（2019-nCoV）感染的肺炎诊疗快速建议指南（标准版）》

中COVID-19 CT相关影像特征比较，找到两者CT影像鉴别方法。 **[结果]** 研究发现早期HAPE与COVID-19胸部CT均表现

为斑片状磨玻璃影，但后者有特征性“铺路石”（网格状小叶内间隔增厚）征象。 **[结果]** 进展期HAPE CT多表现为云絮状

密度增高影，而COVID-19病灶多见平行于胸膜方向发展，部分病灶可见支气管充气征。 **[结果]** 重症爆发期两者CT影像

均可见“白肺”表现，但HAPE右肺多重于左肺。 **[结论]** 因此在HAPE与COVID-19 CT的鉴别诊断中，前者病灶云絮状密度

增高影可作为特征影像征象，而“铺路石征”和“胸膜平行征”可作为后者的特征CT表现。

基金项目语步识别引擎

■ 基金项目非结构化摘要

基于群论的预应力索杆体系刚度解析与形态优化研究

基本信息

批准号	51508089
项目名称	基于群论的预应力索杆体系刚度解析与形态优化研究
项目类别	青年科学基金项目

项目摘要

中文摘要

预应力索杆结构构型新颖、高效，借助体内自平衡预应力形成或改善结构刚度，具有很强的生命力和良好的工程应用前景。针对新型预应力索杆体系设计中的刚度不确定性问题及结构构型与结构性能相互耦合的问题，本项目引入群论方法，开展预应力索杆体系刚度解析及形态优化研究。主要内容包括：（1）研究索杆体系静动不定性及其机理，定性分析机构位移和自应力模态的对称表示，并完善索杆结构预应力稳定性的判定方法；（2）对索杆体系的弹性刚度、几何刚度、结构刚度进行解析，给出评判对称索杆结构稳定性的充分和必要条件；（3）根据等效变换的不可约表示，提出基于群论的结构对称性自动识别方法；（4）分别以预应力和拓扑关系为自变量，并考虑结构性能、经济性和对称性，采用两种独立的约束优化模型，提出基于蚁群算法的预应力索杆体系形态优化方法。本项目一方面丰富和发展索杆体系研究的理论与方法；另一方面为开发大型预应力索杆体系提供新的技术手段与参考。

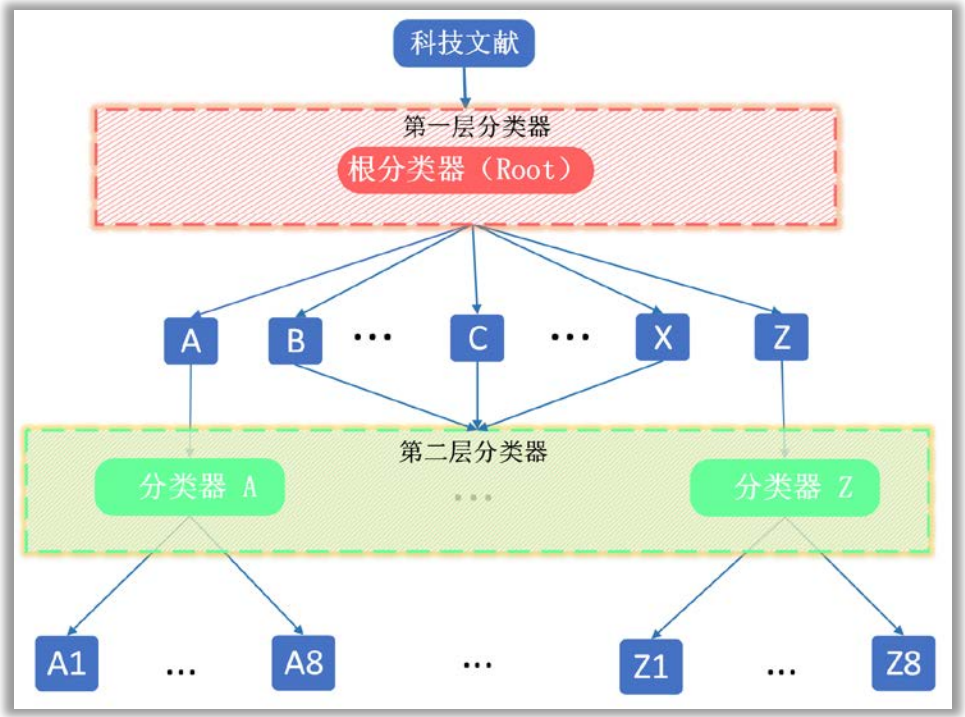
■ 语步识别自动标注结果

[背景及问题] 预应力索杆结构构型新颖、高效，借助体内自平衡预应力形成或改善结构刚度，具有很强的生命力和良好的工程应用前景。 **[目标及任务]** 针对新型预应力索杆体系设计中的刚度不确定性问题及结构构型与结构性能相互耦合的问题，本项目引入群论方法，开展预应力索杆体系刚度解析及形态优化研究。 **[方法内容]** 主要内容包括：（1）研究索杆体系静动不定性及其机理，定性分析机构位移和自应力模态的对称表示，并完善索杆结构预应力稳定性的判定方法；（2）对索杆体系的弹性刚度、几何刚度、结构刚度进行解析，给出评判对称索杆结构稳定性的充分和必要条件；（3）根据等效变换的不可约表示，提出基于群论的结构对称性自动识别方法；（4）分别以预应力和拓扑关系为自变量，并考虑结构性能、经济性和对称性，采用两种独立的约束优化模型，提出基于蚁群算法的预应力索杆体系形态优化方法。 **[价值意义]** 本项目一方面丰富和发展索杆体系研究的理论与方法；另一方面为开发大型预应力索杆体系提供新的技术手段与参考。

SciAIEngine功能

- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

支撑中图法千级类目的多层级分类模型



基于BERT 的两层分类模型

- 医学领域3200篇测试集结果（共计112个类目）

评估指标	单层分类模型	两层分类模型
Precision	0.7826	**0.8051**
Recall	0.7189	**0.7578**
F1-score	0.7184	**0.7623**

实验表明：BERT两层分类模型相比单层分类方法，效果提升4.39%

基于微服务及批处理方式的模型预测性能优化

- 单篇文档预测效率分析

	原始预测	微服务方式
单层分类	32.2899 s	**0.2847 s**
双层分类	65.1954 s	**0.5143 s**

- 600篇文档预测效率分析

	原始预测	微服务方式	微服务 + 批处理方式
单层分类	41.3890 s	14.2058 s	**14.2058 s**
双层分类	114 mins	8 mins	**32.9115 s**

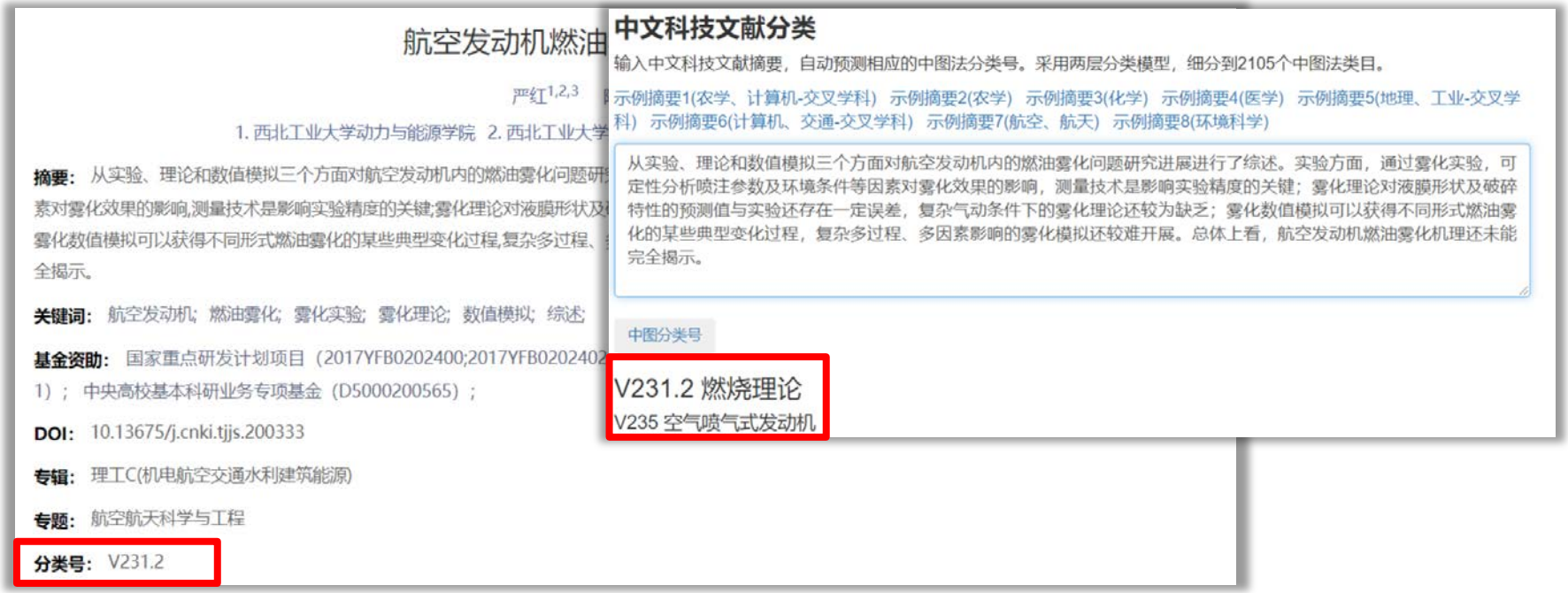
中文科技文献分类引擎

- 基于180万篇科技文献进行训练，实现多达2105个类目的深层次细粒度分类

类目	数据量	类别数	类目	数据量	类别数
医学（R）	448,534	482	一般工业技术（TB）	32,017	28
农学（S）	176,659	276	石油、天然气工业（TE）	32,241	42
计算机、自动化（TP）	181,974	40	轻工业、手工业（TS）	48,523	47
数理化科学（O）	109,920	108	金属学与金属工艺（TG）	62,891	94
天文、地球科学（P）	104,002	147	建筑科学（TU）	61,865	95
生物科学（Q）	67,109	116	机械、仪表工业（TH）	32,970	46
交通运输（U）	41,632	79	能源与动力工程（TK）	25,136	22
航天、航空（V）	27,652	49	水利工程（TV）	15,885	32
环境科学（X）	90,621	78	矿业工程（TD）	10,956	21
电工技术（TM）	59,720	69	原子能工程（TL）	9,744	18
化学工业（TQ）	57,747	88	冶金工业（TF）	8,257	12
无线电电子学、电信技术（TN）	78,253	101	武器工业（TJ）	6,834	15
合计				**1,791,142**	**2105**

中文科技文献分类引擎

- 原始文献分类号
- 分类引擎自动分类结果



SciAIEngine功能

- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

关键词识别

- 自动识别科技文献中的关键词，以先进的关键词识别模型为基础

模型名称	模型介绍
BERT_SoftMax	在预训练语言模型BERT之上添加一个SoftMax分类层，并对BERT进行微调。
BERT_POS_SoftMax	在BERT_SoftMax的基础之上，使用Hanlp进行词性标注，并将词性特征融合到BERT模型当中进行训练。
BERT_Lexicon_SoftMax	在BERT_SoftMax的基础之上，我们构建了医学领域词典，并将词典特征融合到BERT模型中进行训练。
BERT_CRF	在BERT之上使用CRF捕获标签之间的序列特征
BERT_Span	将关键词抽取定义为Span Prediction问题，预测关键短语的开始位置和结束位置。

中文科技文献关键词识别引擎

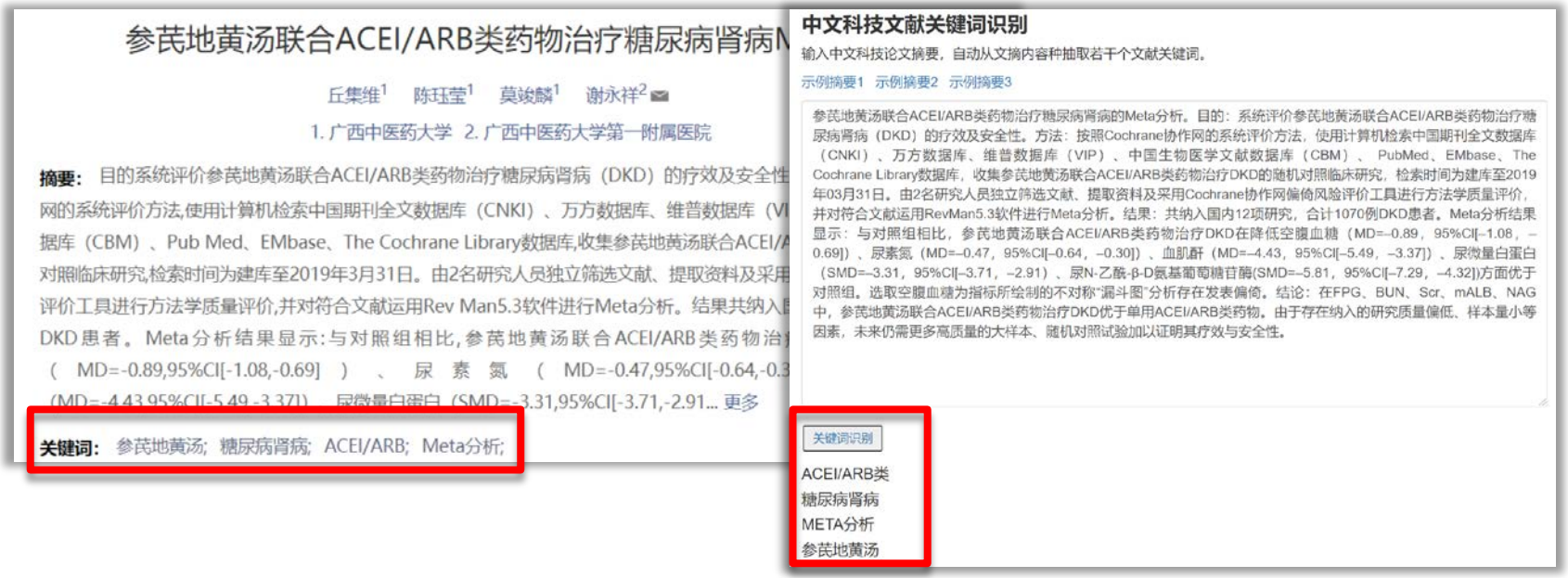
- 基于110万篇中文科技文献进行训练，采用BERT_Lexicon_SoftMax 模型提供关键词识别服务。

Abstract	Labels
['meo', '卫', '星', '内', '部', '充', '电', '环', '境', '及', '典', '型', '材', '料', '充', '电', '特', '征', '分', '析', '。']	['B', 'I', 'I', 'B', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
['阿', '司', '匹', '林', '联', '合', '替', '格', '瑞', '洛', '致', '严', '重', '下', '消', '化', '道', '出', '血', '。']	['B', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B', 'I', 'I', 'I', 'I', 'I', 'O']

- 训练语料示例

中文科技文献关键词识别引擎

- 原始文献关键词
- 引擎自动识别的关键词



SciAIEngine功能

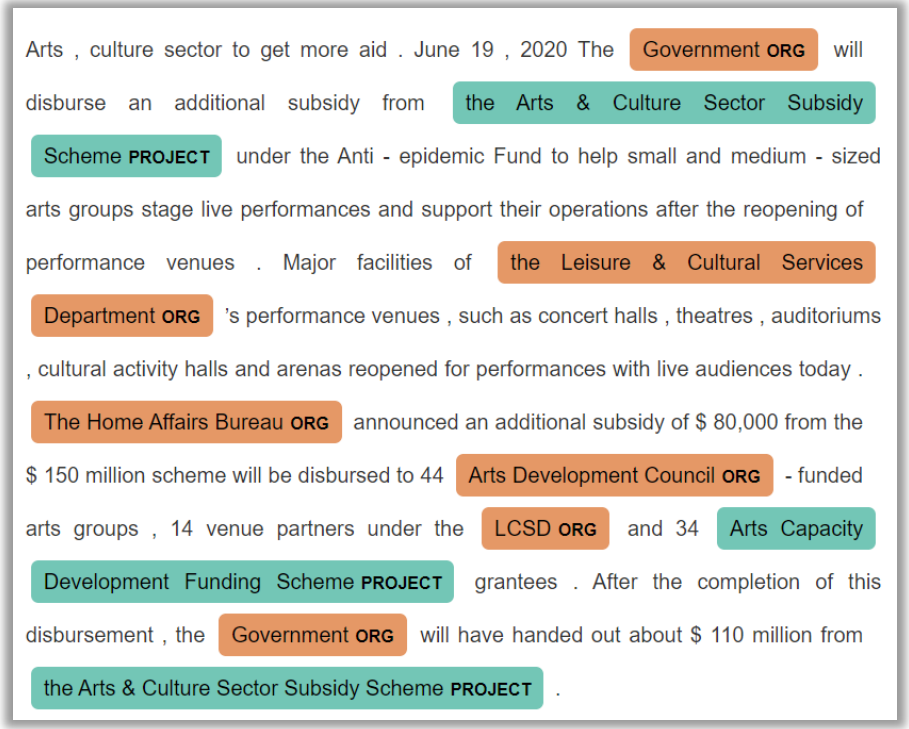
- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

中文通用领域命名实体识别引擎

- 自动识别中文科技文献中的通用命名实体，包含实体类别人名、地名、机构名等

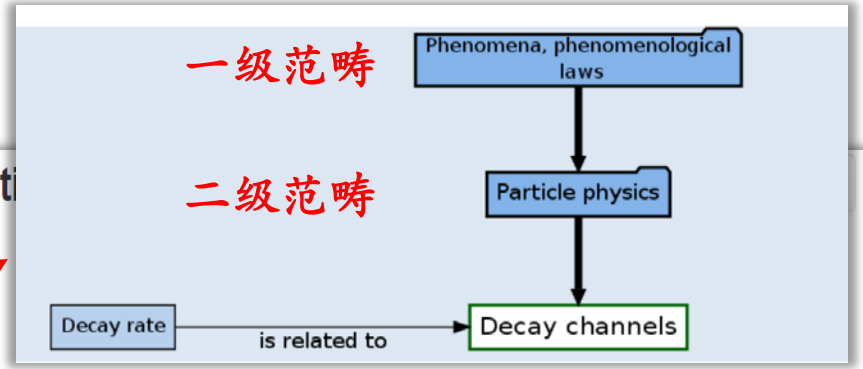
英文通用领域命名实体识别引擎

- 自动识别英文科技文献中的通用命名实体，包含实体类别人名、地名、机构名等



英文物理学领域命名实体识别引擎

- 从物理学本体ScienceWISE获取摘要和术语本体、一级范畴、二级范畴作为训练语料



Search for Heavy Isosinglet Neutrino in e+e- Annihilati

L3 Collaboration

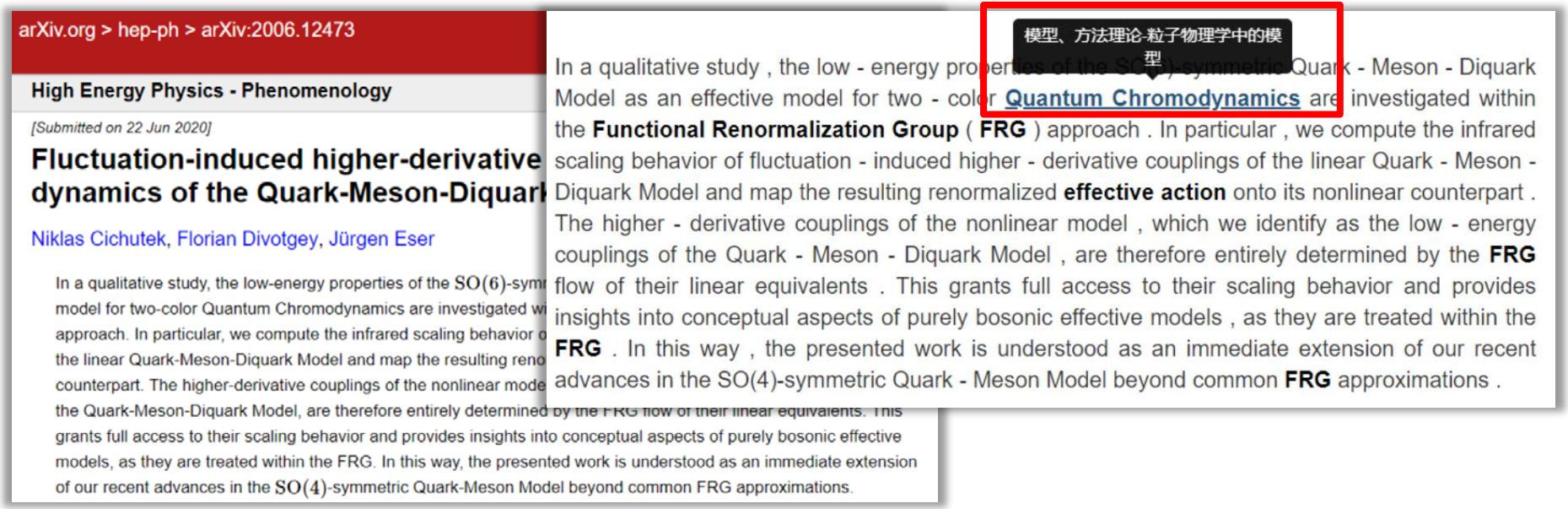
High Energy Physics - Experiment

术语本体

We report on a search for the first generation heavy neutrino that is an isosinglet under the standard SU(2)_L gauge group. The data collected with the L3 detector at center-of-mass energies between 130 GeV and 208 GeV are used.The decay channel N_e --> eW is investigated and no evidence is found for a heavy neutrino, N_e, in a mass range between 80 GeV and 205 GeV. Upper limits on the mixing parameter between the heavy and light neutrino are derived.

英文物理学领域命名实体识别引擎

- 英文物理学摘要
- 实体识别自动标注结果



arXiv.org > hep-ph > arXiv:2006.12473

High Energy Physics - Phenomenology

[Submitted on 22 Jun 2020]

Fluctuation-induced higher-derivative dynamics of the Quark-Meson-Diquark

Niklas Cichutek, Florian Divotgey, Jürgen Eser

In a qualitative study, the low-energy properties of the $SO(6)$-symm model for two-color Quantum Chromodynamics are investigated wi approach. In particular, we compute the infrared scaling behavior o the linear Quark-Meson-Diquark Model and map the resulting reno counterpart. The higher-derivative couplings of the nonlinear mode the Quark-Meson-Diquark Model, are therefore entirely determined by the FRG flow of their linear equivalents. This grants full access to their scaling behavior and provides insights into conceptual aspects of purely bosonic effective models, as they are treated within the FRG. In this way, the presented work is understood as an immediate extension of our recent advances in the $SO(4)$-symmetric Quark-Meson Model beyond common FRG approximations.

模型、方法理论-粒子物理学中的模型

In a qualitative study , the low - energy properties of the SO(4)-symmetric Quark - Meson - Diquark Model as an effective model for two - color Quantum Chromodynamics are investigated within the **Functional Renormalization Group** (**FRG**) approach . In particular , we compute the infrared scaling behavior of fluctuation - induced higher - derivative couplings of the linear Quark - Meson - Diquark Model and map the resulting renormalized **effective action** onto its nonlinear counterpart . The higher - derivative couplings of the nonlinear model , which we identify as the low - energy couplings of the Quark - Meson - Diquark Model , are therefore entirely determined by the **FRG** flow of their linear equivalents . This grants full access to their scaling behavior and provides insights into conceptual aspects of purely bosonic effective models , as they are treated within the **FRG** . In this way , the presented work is understood as an immediate extension of our recent advances in the SO(4)-symmetric Quark - Meson Model beyond common **FRG** approximations .

SciAIEngine功能

- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

科技文献概念定义句识别引擎

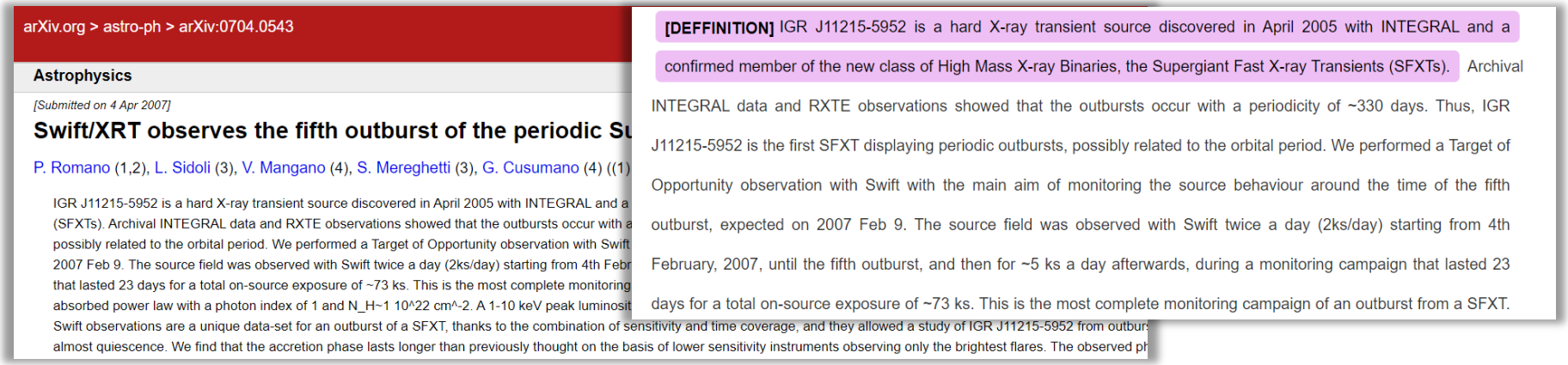
- 基于54万条概念定义句进行训练，自动识别科技文献中的概念定义句子

Label	Sentence
1	Keynsham railway station is a railway station serving the town of Keynsham in Bath and North East Somerset, England.
0	Situated 6 miles north of Driffield town centre and lying on the B1249 between Driffield and Foxholes.
1	A Coulter counter is an apparatus for cell counting and counting and sizing particles suspended in electrolytes.
0	Carl Barc, of the band Ever Since Day One, took over on bass and Dave Karcich, formerly of Spring Heeled Jack, took over on drums.

- 训练语料示例

科技文献概念定义句识别引擎

- 英文科技文献摘要
- 概念定义句自动标注结果



arXiv.org > astro-ph > arXiv:0704.0543

Astrophysics

[Submitted on 4 Apr 2007]

Swift/XRT observes the fifth outburst of the periodic Su

P. Romano (1,2), L. Sidoli (3), V. Mangano (4), S. Mereghetti (3), G. Cusumano (4) ((1)

IGR J11215-5952 is a hard X-ray transient source discovered in April 2005 with INTEGRAL and a (SFXTs). Archival INTEGRAL data and RXTE observations showed that the outbursts occur with a possibly related to the orbital period. We performed a Target of Opportunity observation with Swift 2007 Feb 9. The source field was observed with Swift twice a day (2ks/day) starting from 4th Febr that lasted 23 days for a total on-source exposure of ~73 ks. This is the most complete monitoring absorbed power law with a photon index of 1 and N_H~1 10^22 cm^-2. A 1-10 keV peak luminosit Swift observations are a unique data-set for an outburst of a SFXT, thanks to the combination of sensitivity and time coverage, and they allowed a study of IGR J11215-5952 from outburs almost quiescence. We find that the accretion phase lasts longer than previously thought on the basis of lower sensitivity instruments observing only the brightest flares. The observed ph

[DEFFINITION] IGR J11215-5952 is a hard X-ray transient source discovered in April 2005 with INTEGRAL and a confirmed member of the new class of High Mass X-ray Binaries, the Supergiant Fast X-ray Transients (SFXTs). Archival INTEGRAL data and RXTE observations showed that the outbursts occur with a periodicity of ~330 days. Thus, IGR J11215-5952 is the first SFXT displaying periodic outbursts, possibly related to the orbital period. We performed a Target of Opportunity observation with Swift with the main aim of monitoring the source behaviour around the time of the fifth outburst, expected on 2007 Feb 9. The source field was observed with Swift twice a day (2ks/day) starting from 4th February, 2007, until the fifth outburst, and then for ~5 ks a day afterwards, during a monitoring campaign that lasted 23 days for a total on-source exposure of ~73 ks. This is the most complete monitoring campaign of an outburst from a SFXT.

SciAIEngine功能

- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

中文科技文献标题生成引擎

- 基于50万篇中文文献进行训练，自动生成表达科技文献内容的文献标题

Label	Abstract
非均相复合驱数值模拟方法研究与应用	非均相复合驱是一项应用于聚合物驱后油藏进一步提高采收率的化学驱方法,其主要的驱替剂为预交联凝胶颗粒B-PPG、聚合物和表面活性剂……
对机载相控阵雷达STAP 技术的旁瓣干扰	空时自适应处理(STAP)是一种有效的抗干扰技术。介绍机载相控阵雷达STAP 技术检测动目标的基本原理,提出了基于灵巧噪声的旁瓣干扰方法……
低影响开发(LID)生物滞留技术研究进展	随着我国城市化进程的加快,由城市下垫面改变和降水径流引发的环境问题日益严重,作为低影响开发措施之一,生物滞留技术对于消纳、净化降水径流具有重要作用……
预设路径模型及其在认知心理学研究中的应用	预设路径模型(Fixed-links modeling)是在结构方程模型框架下发展出的用于分析心理学实验数据的统计模型。该类模型的主要特征是根据前期理论基础和实验设计……

- 训练语料示例

中文科技文献标题生成引擎

- 原始科技文献标题

- 引擎自动生成的标题

SciAIEngine功能

- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

中文科技文献审稿人推荐引擎

- 基于78万篇科技文献和4万多位作者进行深度学习训练，自动推荐适合评审某篇论文的审稿专家

Label	Abstract
龚振平（东北农业大学农学院）	氮素与根瘤固氮结合可达高产,两者矛盾对根瘤固氮产生不利影响。氮素影响根瘤固氮作用机制仍不明确。文章在现有研究成果基础上,总结氮素与大豆根瘤固氮关系研究……
董辉（东北大学）	烧结余热罐式回收系统是针对于传统烧结余热回收系统的不足,借鉴干熄焦(CDQ)提出的一种变革性烧结余热回收系统,其具有余热回收率较高、漏风率低等优点……
刘雷（哈尔滨工业大学能源科学与工程学院）	为了研究某半埋入式S弯进气道出口畸变对其后风扇级性能的影响,分别将其原型及优化后模型与风扇级对接进行进气道加风扇级全流道数值研究……
刘振军（重庆大学）	建立了湿式双离合器自动变速器(dual clutch transmission, DCT)换挡过程系统动力学模型,针对湿式双离合器系统的高度非线性,难以建立精确数学模型等特点……

- 训练语料示例

中文科技文献审稿人推荐引擎

- 中文科技文献

- 引擎自动推荐的审稿人

SciAIEngine功能

- 语步识别
- 科技文献分类
- 关键词识别
- 命名实体识别
- 概念定义句识别
- 文本标题生成
- 审稿人推荐
- 投稿期刊推荐

中文科技文献投稿期刊推荐引擎

- 基于270万篇科技文献和1585种科技期刊进行训练，自动推荐适合的投稿科技期刊

Label	Abstract
刊名:高原气象 ISSN:1000-0534	选取江苏和内蒙古分别作为中国沿海滩涂与内陆复杂高原山地的典型地形代表,通过中尺度模式WRF V3.3.1两种不同边界层参数化方案(YSU/MRF)的对比检验……
刊名:玉米科学 ISSN:1005-0906	选取中国知网(CNKI)数据库文献作为数据来源,应用文献计量学的方法,分析国内玉米单倍体育种的年度发文量、主要研究机构、高被引文章以及研究方向……
刊名:中国安全生产科学技术	为了采取合理的瓦斯抽采技术,实现矿井安全高效开采,对朱集煤矿13-1煤层实行高位钻孔以及上隅角采空区埋管作为抽采瓦斯实验方案。通过对现场工程实践……
刊名:稀有金属 ISSN:0258-7076	用磁控溅射法在单晶硅和聚酰亚胺衬底上制备了恒定调制比(η=1)、调制周期λ=10~100 nm的Cu/Mo纳米多层膜，运用XRD，HRTEM，EDX，AFM，单轴……

- 训练语料示例

中文科技文献投稿期刊推荐引擎

- 中文科技文献
- 引擎自动推荐的投稿期刊

中国草地学报，2020年05期 第1-7页 北大核心

发根农杆菌介导的箭筈豌豆毛状根遗传转化体系的建立

梅错　刘志鹏

草地农业生态系统国家重点实验室/兰州大学草地农业科技学院

摘要： 首次建立了箭筈豌豆的毛状根遗传转化体系,利用ARqua1和K599两种发根农杆菌成功诱导出阳性转基因毛状根,其更适合诱导箭筈豌豆毛状根的是ARqua1菌株。在无菌和非无菌条件下,ARqua1菌株的阳性毛状根诱导效率分别为81.7%、35.4%,而K599菌株的阳性毛状根诱导效率分别为70.0%、33.3%。PCR检测和GUS染色结果表明,GUS基因在毛状根中稳定表达,毛状根遗传转化体系可应用于箭筈豌豆分子和蛋白水平的初步研究。

关键词： 箭筈豌豆；发根农杆菌；毛状根；遗传转化;

基金资助： 国家自然科学基金优秀青年基金项目“牧草育种学”(31722055)；国家重点研发计划项目“青藏高原退化恢复的主要物源制约因子及其应用技术研发”(2019YFC0507700)；甘肃省科技重大专项项目(19ZD2NA002);

DOI： 10.16742/j.zgcdxb.20190305

专辑： 农业科技

专题： 农作物

分类号： S542.9

中文科技文献投稿期刊推荐

输入中文科技文献摘要文本内容，自动推荐若干与该论文相关的投稿期刊。

基于BERT Fine-tuning完成模型微调。训练语料涵盖全领域270万中文摘要数据，候选期刊1585种。

示例摘要1　示例摘要2　示例摘要3　示例摘要4

首次建立了箭筈豌豆的毛状根遗传转化体系,利用ARqua1和K599两种发根农杆菌成功诱导出阳性转基因毛状根,其中更适合诱导箭筈豌豆毛状根的是ARqua1菌株。在无菌和非无菌条件下,ARqua1菌株的阳性毛状根诱导效率分别为81.7%、35.4%,而K599菌株的阳性毛状根诱导效率分别为70.0%、33.3%。PCR检测和GUS染色结果表明,GUS基因在毛状根中稳定表达,毛状根遗传转化体系可应用于箭筈豌豆分子和蛋白水平的初步研究。

投稿期刊推荐

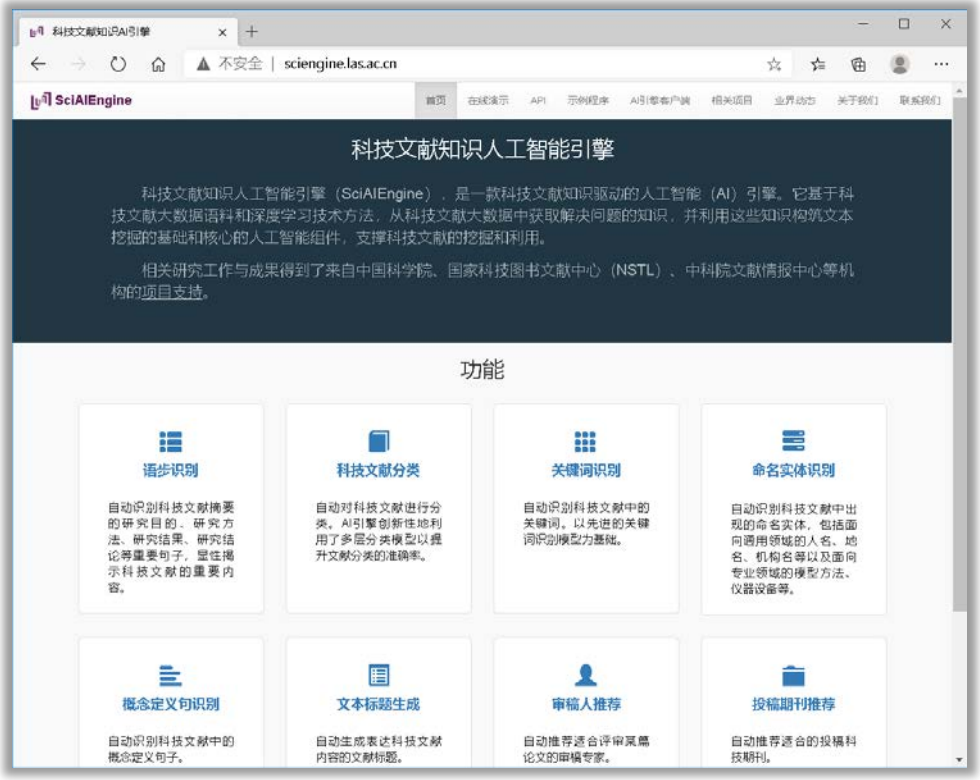
刊名:生物技术通报 ISSN:1002-5464
刊名:中国草地学报 ISSN:1673-5021
刊名:广东农业科学 ISSN:1004-874X
刊名:作物杂志 ISSN:1001-7283
刊名:江苏农业科学 ISSN:1002-1302
刊名:草业学报 ISSN:1004-5759
刊名:华北农学报 ISSN:1000-7091
刊名:植物生理学报 ISSN:2095-1108
刊名:河南农业科学 ISSN:1004-3268
刊名:湖北农业科学 ISSN:0439-8114

SciAIEngine：智慧服务能力提升的实践

- SciAIEngine是什么？
- SciAIEngine解决什么问题?
- SciAIEngine怎样用？
- SciAIEngine都在哪里用了？

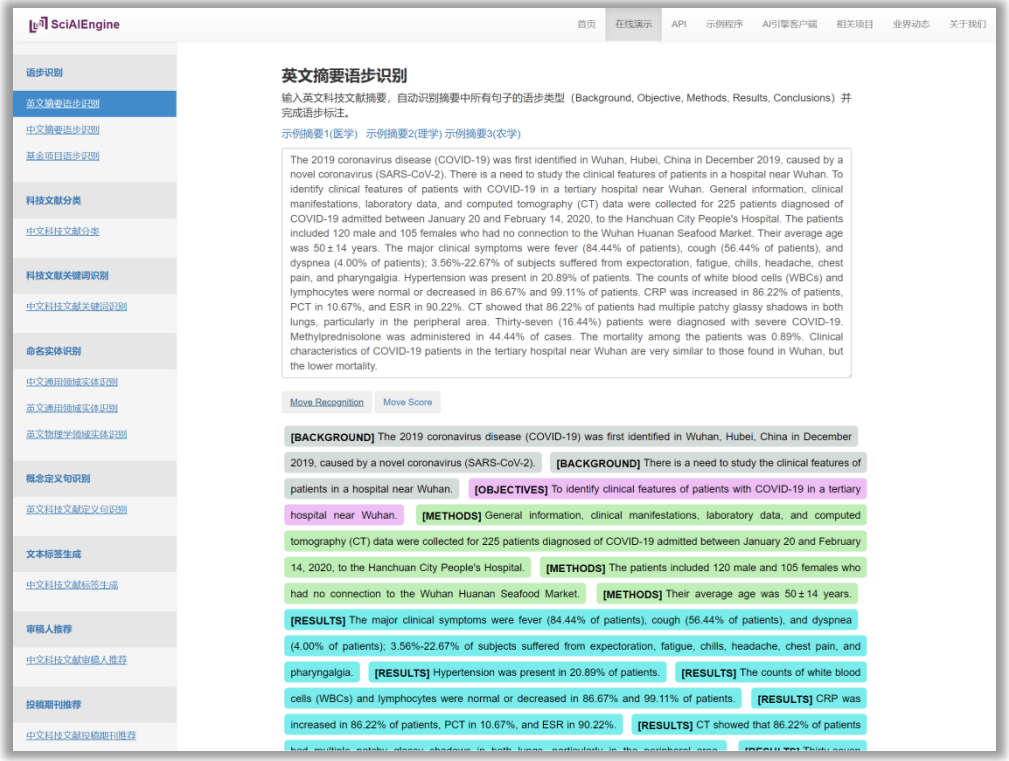
科技文献知识人工智能引擎（SciAIEngine）

http://sciengine.las.ac.cn/



在线表单提交

- 用户可以通过浏览器在线提交需要处理的文献，实时查看AI引擎自动标注的输出结果



HTTP API接口调用

- 用户可以通过GET或POST方式提交待处理的文档，获取AI引擎自动标注的输出结果

GET方式（示例程序）

支持单篇文档，url为http://sciengine.las.ac.cn/move_recognition_en，传入英文摘要文本，将以json格式返回该摘要的语步标注结果。

	格式	示例
请求URL	/move_recognition_en	http://sciengine.las.ac.cn/move_recognition_en
请求参数(JSON)	"data": 摘要文本 "token": 验证码数字	{"data":"The aim of this paper is to study...","token":99999}
浏览器参数访问	/move_recognition_en?data= &token=	http://sciengine.las.ac.cn/move_recognition_en?data=The aim of this paper is to study...&token=99999
成功请求返回数据(JSON)	"语步标签":[句子列表]	{ "Background":[{'The prospect of using...':0}], "Objective": [{'Here, we study the...':1}], "Methods": [{'We consider a situation...':2}], "Results": [{'We show that...':3}], "Conclusions": [{'We find that ...':4},{'This offers the possibility...':5}] }
错误信息返回数据(JSON)	"info":错误信息	{ "info": "Server not available!" } { "info": "Token incorrect!" }

POST方式（示例程序）

支持多篇文档，url为http://sciengine.las.ac.cn/move_recognition_en，将多个摘要文档以list类型上传，将以json格式将多篇摘要标注结果放在列表中返回，其中每篇摘要标注结果格式与上述GET方式一致。

	格式	示例
请求URL	/move_recognition_en	http://sciengine.las.ac.cn/move_recognition_en
请求参数(JSON)	"data": 摘要文本列表 "token": 验证码数字	{ "data":["The aim of this paper is to study...", "The aim of this paper is to study...",],"token":99999 }
成功请求返回数据(JSON)	"results": 结果列表	{ "results":[{ "Background":[...], "Objective": [...], "Methods": [...], "Results": [...], "Conclusions": [...] }, { "Background":[...], "Objective": [...], "Methods": [...], "Results": [...], "Conclusions": [...] }] }
错误信息返回数据(JSON)	"info":错误信息	{ "info": "Server not available!" } { "info": "Token incorrect!" }

下载和重用AI引擎提供的Python示例程序

- 用户可从AI引擎网站上下载相应的示例程序，替换要处理的数据文件，即可调用AI引擎来进行自动标注

示例程序

我们针对平台的API接口，提供相应的Python示例程序。

英文摘要语步识别

相关程序打包下载：move_recognition_en.zip

文件说明：

- move_recognition_en_get.py：GET方式执行文件。
- move_recognition_en_post.py：POST方式执行文件。
- input_en.txt：示例文件，可输入多篇非结构化英文摘要文本，每条摘要为一行。
- ReadMe.txt：说明文件。

使用说明：

- 程序运行环境：Python>=3.5 requests>=2.22.0
- 运行程序前，请先确认您已注册获取Token验证码，注册地址。
- GET方式：使用编辑器打开move_recognition_en_get.py文件，将Token验证码以及需要进行语步识别的英文摘要文本输入到程序中相应位置。运行move_recognition_en_get.py，直接打印语步标注结果。
- POST方式：使用编辑器打开move_recognition_en_post.py文件，将Token验证码输入到程序中相应位置。在程序中设置好输入文件和输出文件的路径，其中输入文件的格式为每篇摘要一行。运行move_recognition_en_post.py，程序读取输入文件，将语步识别结果写入在输出文件中。

利用客户端程序

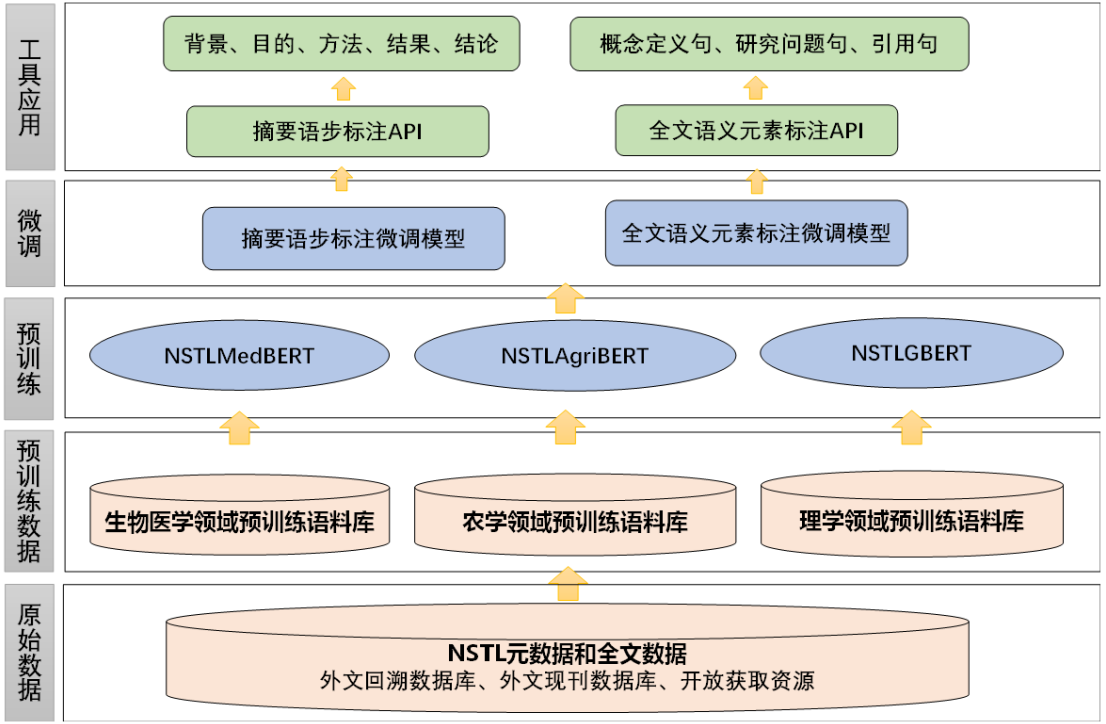
- 用户也可通过下载安装客户端程序，接使用语步识别、文献分类和关键词识别等功能

SciAIEngine：智慧服务能力提升的实践

- SciAIEngine是什么？
- SciAIEngine解决什么问题?
- SciAIEngine怎样用？
- SciAIEngine都在哪里用了？

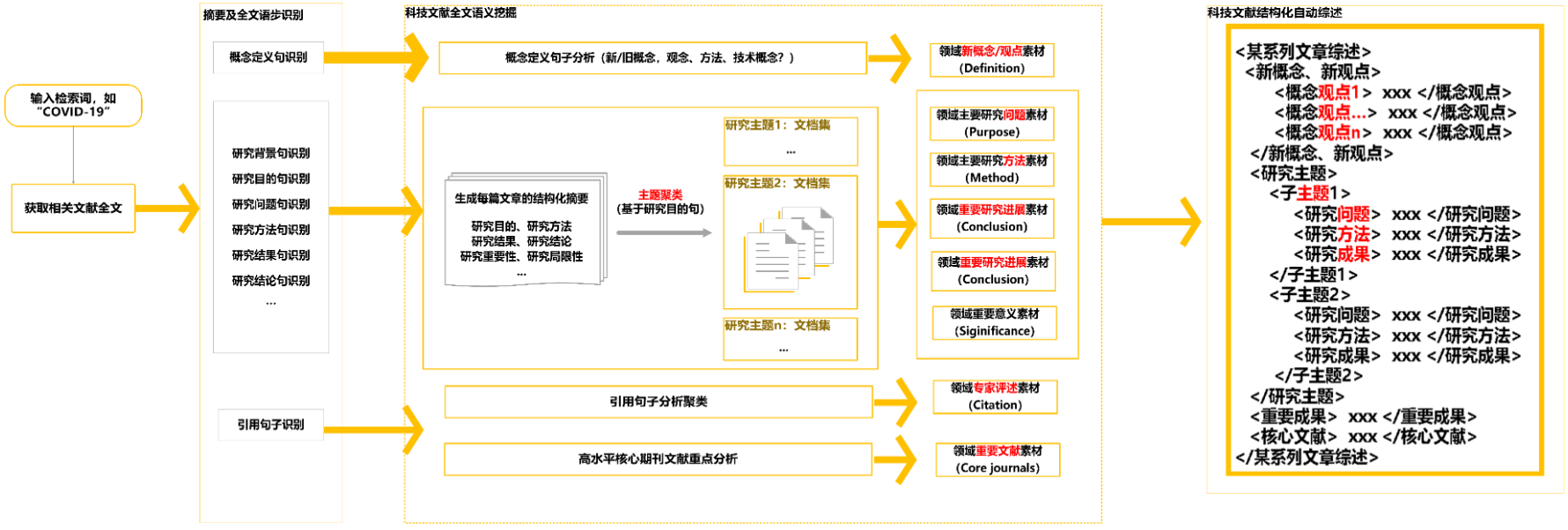
人工智能（AI）引擎推广及示范应用

- NSTL Science watch
  - NSTL“下一代开放知识服务平台总体设计及关键技术研发”专项
  - 承担英文科技论文语步标注工具构建工作
  - 承担科技文献预训练模型构建工作



人工智能（AI）引擎推广及示范应用

- NSTL
  - 尝试进一步实现在科技论文结构化综述自动生成中的应用



人工智能（AI）引擎推广及示范应用

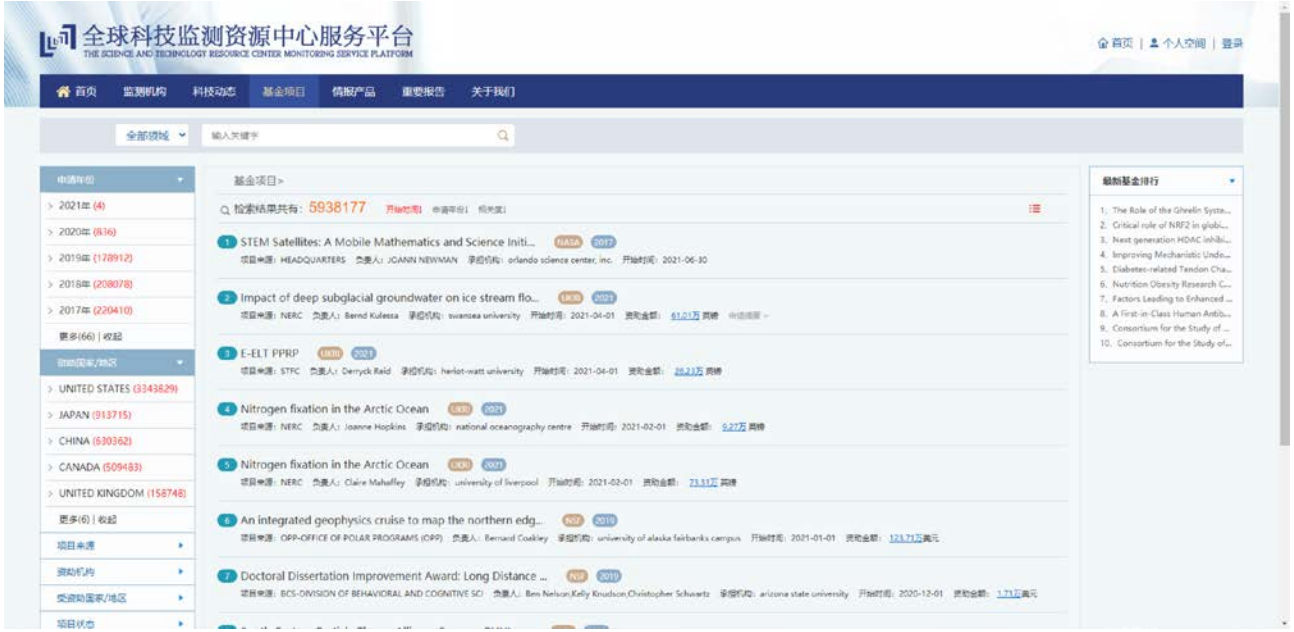
- 中科院文献情报中心
- 学位论文自动分类应用

Id	题名	中图法分类号	摘要
151195	超快激光数控机床控制系统的研究与开发	['TP273 自动控制、自动控制系统', '']	激光加工技术具有效率高、材料消耗低、环保等诸多优点，自诞生之日起就引起人们的高度关注，为此，用于激光加工的数控机床的开发成为数控机床发展的
151196	智能化数控系统体系结构及关键技术研究与实现	['TP273 自动控制、自动控制系统', 'TP393 计算机网络	随着制造技术的迅猛发展，以智能制造所引领的产业变革即将进入重要的发展阶段。为了抢抓新一轮科技革命的发展先机，工业4.0、工业互联网战略、《中
151197	面向宫颈癌辅助诊断的深度学习算法模型研究与实现	['TP391 信息处理', 'TP391.4 模式识别与装置']	宫颈癌是严重危害妇女健康的恶性肿瘤之一，早期筛查可有效检测出宫颈癌及癌前病变，人工手段的细胞学检测方法假阴性率高、准确率低。基于深度学习的
3094	AB类音频DAC设计	['TP33 电子数字计算机', 'TP391 信息处理']	摘要 随着智能手机、平板电脑等便携式移动终端的普及，作为这些设备所必须的音频编码器（音频模数转换器）、音频解码器（音频数模转换器）的需求也
10558	面向微博的动态短文本检索技术研究	['TP391 信息处理', 'TP391.1 文字信息处理']	随着 Web2.0 的快速发展,互联网上积累了海量的用户产生内容,如微博、论坛、 贴吧、评论等。这类信息资源具有两个显著的特点:(1)文档具有时间动态性
10727	结构稀疏学习及其在图像检索中的应用研究	['TP391 信息处理', 'TP391.4 模式识别与装置']	在数字信号处理和机器学习领域中，稀疏表示是一个基础且重要的研究问题，并在大量的实际应用中均表现出优异的性能。稀疏学习是一类基于稀疏表示的
20980	面向图像分类的深度残差网络优化结构研究	['TP391 信息处理', 'TP391.4 模式识别与装置']	近年来，随着计算机性能的提升，以及网络多媒体数据爆炸式的增长，海量多媒体数据的处理受到广泛关注。在计算机视觉领域，由于深度学习方法的发展，
20514	一种面向容器环境的监测系统设计与实现	['TP311.5 软件工程', 'TP393 计算机网络']	容器（Docker）的核心思想是OS复用，为应用提供沙箱运行环境，具有资源供给快速轻量的优势，因此该技术已经被广泛用来应对互联网应用资源突变的需
20895	PWELL注入对0.11微米产品良率影响研究	['TP391 信息处理', 'TP391.7 机器辅助技术']	在集成电路制造工艺中，因使用的硅基体是半导体材料，为增强其导电性，目前比较成熟的工艺都是采用离子注入的方式。离子注入的应用非常广泛，如形
20521	一种融合草图交互的视频内容可视分析方法	['TP391 信息处理', 'TP391.4 模式识别与装置']	随着数码设备的普及和互联网技术的发展，视频资源数量呈现海量式增长。人们不再满足于传统单一线性的视频信息利用方式，这对有效地检索信息、快速
20901	基于在线产品评论的观点挖掘系统研究	['TP391 信息处理', 'TP391.1 文字信息处理']	随着电子商务的发展，互联网上存在着海量的网购用户对产品的评论信息。这些评论信息中蕴含着消费者使用产品后的主观意见、情感和态度等，因而具有
20496	数据中心虚拟网络拥塞控制机制研究	['TP393 计算机网络', '']	云数据中心环境中，虚拟机已取代物理机作为网络的基本单位及租赁的基本单位。虚拟机数量的剧增为网络引入庞大的MAC地址空间。虚拟机任意迁移的需求
20841	基于代码变换的扩展内存接口技术及其应用	['TP316 操作系统', 'TP311.1 程序设计']	扩展内存技术是近年来的一个研究热点，广泛应用于多种场景之中，如搜索引擎、推荐系统等。学术界和产业界提出了多种扩展内存技术，但是这些扩展内存
20903	SandroidSat-1卫星姿态确定系统研究	['TP21 自动化元件、部件', 'TP391 信息处理']	本文以SandroidSat-1的姿态确定系统设计为背景，从理论上全面的研究了以SandroidSat-1为代表的纳卫星的姿态确定系统，建立了姿态敏感器的测量模型
20934	齿脸形状匹配关系研究	['TP391 信息处理', 'TP391.4 模式识别与装置']	牙齿修复是口腔医学研究中的重要课题。在现代口腔修复医学中，选择合适的牙齿齿型以满足患者的生理、心理和社会需要成为齿型选择的重要标准。选择合
20531	多核的超限学习集成方法研究及其在心电房颤预警中的应	['TP391 信息处理', 'TP181 自动推理、机器学习']	随着可穿戴医疗设备的发展和广泛使用，大量的人体生理数据被采集并保存下来，对这些数据进行快速有效的分析将面临很多挑战。大规模心电数据的分
20658	面向大数据的文本分类系统的设计与实现	['TP391 信息处理', 'TP391.1 文字信息处理']	摘要信息技术的发展使得互联网数据呈爆炸式增长，互联网数据中有大量的数据以文本的形式存在，怎样从海量的数据中快速定位到自己所需的信息已成为一
20532	海计算机运行时系统设计与实现	['TP393 计算机网络', 'TP311.5 软件工程']	随着物联网技术的发展，越来越多的智能设备将会走进我们的生活。由于不同行业针对特定应用设计硬件结构，根据特定业务搭建相应的平台。导致目前物联
20597	多源异构数据转换系统的设计与实现	['TP391 信息处理', 'TP311.1 程序设计']	随着大数据的蓬勃发展，数据平台建设已经被各个企业提上日程。然而，由于企业发展过程中，不同应用采取的数据存储系统不尽相同，但历史数据的价值
20688	鲁棒的视觉单目标跟踪算法研究	['TP391.4 模式识别与装置', 'TP391 信息处理']	视觉跟踪是计算机视觉中一个非常重要并且十分活跃的研究方向。它在人机交互、行为分析、车辆导航以及智能监控等领域有着广阔的应用前景，研究视觉跟
20598	面向海量图数据的分布式动无关异步引擎研究	['TP391 信息处理', 'TP311.1 程序设计']	随着大数据时代的到来，图算法越来越受到人们的重视，其在社交网络分析、舆情分析、推荐系统等领域中作用越来越大，因此出现了许多分布式图计算框
20661	龙芯2H平台网络转发性能优化	['TP393 计算机网络', '']	随着网络技术的普及，网络安全的重要性显得越发突出，尤其是在要求完全自主可控的国产处理器中。龙芯系列处理器是我国拥有自主知识产权的高性能处
20501	HS-SWD:基于瓦记录磁盘的混合存储系统	['TP33 电子数字计算机', '']	随着数据爆发式增长，对存储系统的容量提出了巨大挑战。而作为主流存储介质的磁盘，在传统磁记录技术下，其面密度的增长受到限制，迫切需要新的
20536	wifi-zigbee 双模节点通信模式自切换和协作通信技术研	['TP393 计算机网络', '']	物联网凭借其连通互联网世界和真实物理世界的能力，已经在工业、医疗、交通、民生等各个领域获得广泛应用。无线传感器网络作为物联网前端感知物理世
20537	面向复杂依赖关系RPM包的构建方法研究与实现	['TP393 计算机网络', 'TP311.5 软件工程']	Linux系统具有稳定性高、安全性强、开放自由等优点，因而在服务器和嵌入式领域被广泛地使用。RPM是一种Linux软件包管理器，国外的Redhat、Fedora
20662	桌面搜索若干关键技术研究与系统实现	['TP393 计算机网络', 'TP391 信息处理']	随着信息的海量增长，互联网环境以及用户个人设备如PC、手机中的数据量急剧增加。作为应对互联网信息过载的重要方式，网络搜索引擎在用户检索信息
20817	面向高并发视频传输的无线接入技术研究与实现	['TP393 计算机网络', 'TP391 信息处理']	随着信息技术的发展，视频监控系统已经广泛应用于工业、安防、交通、生活等各个领域。视频监控系统的核心技术是视频传输技术，无线视频传输技
20818	面向文本检索的排序学习训练查询选择算法研究	['TP391 信息处理', 'TP391.1 文字信息处理']	文本检索是信息检索技术最重要的应用之一，近数十年来相关学术研究方兴未艾。随着互联网的迅猛发展，传统意义上的单一检索模型，如布尔检索、向量空
20507	面向NFV的高速网络负载均衡系统的设计与实现	['TP393 计算机网络', '']	在传统电信产业中，产品由各式各样的硬件设备承载，而随着新技术和服务的创新加速，硬件设备的生命周期越来越短，服务提供商的网络运营日趋复杂，
20692	服装行业消费者与设计师交互系统研究与实现	['TP391 信息处理', 'TP399 在其他方面的应用']	同质化的困境给国内的很多服鞋业生产商带来了巨大的困扰，而工业4.0的战略报告给国内的服鞋业生产企业带来了新的希望。在详细的研究了国内服鞋企业
20819	基于深度图的三维视频编码优化技术研究	['TP391 信息处理', 'TP37 多媒体技术与多媒体计算机'	随着人们对视觉体验的要求不断提高，三维视频成了当下炙手可热的技术，市场对三维视频的质量要求越发严格。为了满足这种需求，国际电信联盟的视频
20820	大数据环境下特征提取算法与应用研究	['TP391.4 模式识别与装置', 'TP391 信息处理']	随着移动设备与社交网络的急速发展，互联网中的图像数据呈现指数级爆炸增长。然而，海量图像数据中存在着大量的冗余信息和复杂特征，极大
20505	QT开发框架在MIPS平台上的移植与优化	['TP311.5 软件工程', 'TP391 信息处理']	QT开发框架是国内MIPS平台上主流的GUI程序研发工具，由于提供了完善的编程工具和强大的API，因而QT能够加速开发人员的研发工作。在当前的MIPS平台
20601	Linux内核热修复框架的设计和实现	['TP393 计算机网络', '']	随着大数据时代的来临，人们对信息设备的依赖与日俱增。然而，设备系统中存在的漏洞对使用者的信息安全造成了巨大威胁。为了及时修复系统漏洞
20693	面向Cache-Aside模式的内存数据网格关键技术的设计与实	['TP393 计算机网络', 'TP311.1 程序设计']	随着网络技术的发展，数据量和网络流量出现了爆炸性的增长，传统关系型数据库在高并发读写、海量数据存储和系统高扩展性等方面存在明显局限性。内存
20509	CMR-1 3G 信令处理单元的设计与实现	['TP393 计算机网络', '']	对于整个通信系统而言，卫星移动通信系统是地面蜂窝系统的一种必要的补充和延伸，尤其在抢险救灾、应急指挥、海洋渔业等领域发挥重要作用。卫星移动
21403	基于日志的云服务监控与预警系统研究与实现	['TP393 计算机网络', 'TP391 信息处理']	随着互联网和云计算技术的迅猛发展，互联网在线云服务模式已逐渐被用户接受和使用，云服务的能力和规模也急剧膨胀。云服务的分布式特性和高可扩展性
21067	多尺度场景流估计方法	['TP391 信息处理', 'TP391.4 模式识别与装置']	三维空间中的目标和场景的运动称为场景流，它们在二维图像平面的投影形成了运动，这种运动以图像平面亮度模式表现出来的流动就称为光流。场景流估计

分类结果 | 人工评价（两个分类号） | 人工评价（一个分类号）

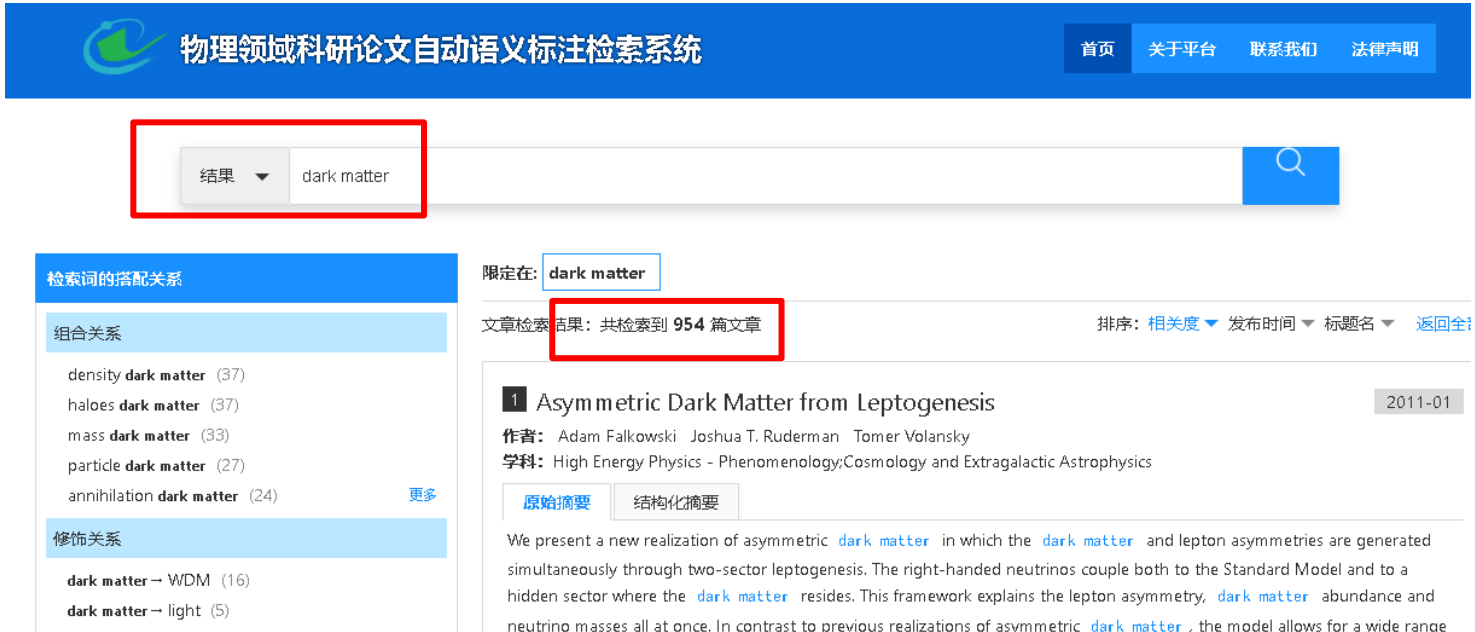
人工智能（AI）引擎推广及示范应用

- 中科院文献情报中心
- 基金项目文本语步识别应用



人工智能（AI）引擎推广及示范应用

- 中科院文献情报中心
- 语义智能检索系统应用



提纲

- 知识获取能力：AI飞速突破的本质
- 科技文献库：图书馆智慧服务的一把钥匙
- SciAIEngine：智慧服务能力提升的思路
- SciAIEngine：智慧服务能力提升的实践
- 下一步的工作

下一步工作安排

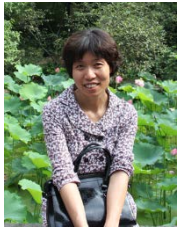
- 语料收集整理工作
  - 进一步收集、整理预训练语料库，尤其是特定领域科技文献全文、书籍、知识库、术语词表等
- 应用推广工作待完善
  - 进一步完善基于科技文献知识资源的AI引擎平台建设，完善相关功能，提升平台运行效率，完善对外服务API接口建设
- 在发布AI引擎之后，做好服务
  - 在12月4日已经发布了引擎
- 实际应用效果待提升
  - 基于科技文献情报工作的实际应用场景，进一步实验、测试模型的应用效果，根据实际情况调整相关研究的改进方向，形成实际可用、有效的工具化的模型

总结

- 以大规模科技文献内容为训练语料的文本挖掘工具!
- 将“科技文献库”转变成为“人工智能引擎”的创新性成果!
- 支撑科技文献挖掘和智慧知识服务的战略性基础设施!
- 得到国家科技图书文献中心（NSTL）和中国科学院文献情报系统大力支持的，中国文献情报机构贡献给全球人工智能时代的重要智慧解决方案!

中国科学院文献情报中心强有力技术团队

张智雄（研究员）
课题负责人



刘筱敏（研究馆员）
引擎推广

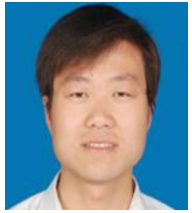
刘小兵（副研究馆员）
数据整理

于改红（馆员）
语义标注

余丽（馆员）
信息抽取

景然（馆员）
数据整理

马娜（馆员）
引擎推广



张敏（馆员）
语义索引

刘熠（博士后）
智能问答

刘欢（博士）
预训练模型
引擎构建

丁良萍（博士）
关键词抽取
实体识别

李婕（博士）
审稿人推荐
期刊推荐

赵旸（博士）
客户端构建
文献分类

李雪思（博士）
定义句识别

王宇飞（博士）
文本标题生成