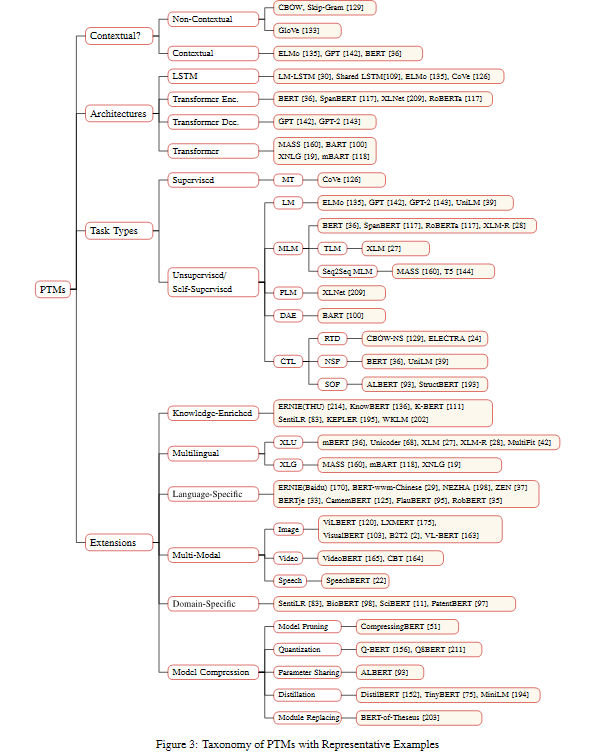
资料来源：

[1] Qiu, Xipeng & Sun, Tianxiang & Xu, Yige & Shao, Yunfan & Dai, Ning & Huang, Xuanjing. (2020). Pre-trained Models for Natural Language Processing: A Survey.

[2] <https://www.jianshu.com/p/17226c9e24b8>

[3] <https://zhuanlan.zhihu.com/p/270009212?utm_source=com.tencent.tim>

**一、预训练模型分类**



（一）按范式分类：

1、第一代：词嵌入模型

例如：word2vec(CBOW/Skip-Gram)，paragraph vector，Skip-thought vectors, Context2Vec

2、第二代：其他深度模型

例如：LSTM、ELMo，GPT，BERT

（二）任务区分

1、监督学习（Supervised learning）

最常见的是CoVe

2、无监督学习

3、自监督学习

（1）基于上下文：

语言模型LM；去噪自编码器DAE；掩码语言模型MLM；排列语言模型PLM

（2）基于对比：

最大化互信息DIM；替换token检测RTD； 预测下一句NSP；句子顺序预测SOP

**二、一些开源预训练模型的github地址**

word2vec:

https://github.com/tmikolov/word2vec

GloVe:

https://nlp.stanford.edu/projects/glove

FastText:

https://github.com/facebookresearch/fastText

Transformers:

https://github.com/huggingface/transformers

Fairseq:

https://github.com/pytorch/fairseq

Flair:

https://github.com/flairNLP/flair

AllenNLP:

https://github.com/allenai/allennlp

FastNLP:

https://github.com/fastnlp/fastNLP

Chinese-BERT:

https://github.com/ymcui/Chinese-BERT-wwm

BERT:

https://github.com/google-research/bert

RoBERTa:

https://github.com/pytorch/fairseq/tree/master/examples/roberta

XLNet:

https://github.com/zihangdai/xlnet/

ALBERT:

https://github.com/google-research/ALBERT

T5:

https://github.com/google-research/text-to-text-transfer-transformer

ERNIE (Baidu):

https://github.com/PaddlePaddle/ERNIE

**三、预训练模型与知识图谱相结合的研究进展**

1、ERNIE: Enhanced Language Representation with Informative Entities

这篇论文来自于清华刘知远老师和华为刘群老师，已被ACL2019所录取，是较早的考虑将知识引入预训练模型的论文。

该论文主要利用了从知识库中提出的高信息量的实体信息，通过特殊的语义融合模块，来增强文本中对应的表示。

2、K-BERT: Enabling Language Representation with Knowledge Graph

这篇论文来自于北大和腾讯，已被AAAI2020所录取，是较早的考虑将知识图谱中的边关系引入预训练模型的论文。

该论文主要通过修改Transformer中的attention机制，通过特殊的mask方法将知识图谱中的相关边考虑到编码过程中，进而增强预训练模型的效果。

3、KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

这篇论文来源于清华和Mila实验室，其主要关注于如何使用BERT增强知识图谱embedding，并帮助增强对应的表示。

该论文主要通过添加类似于TransE的预训练机制来增强对应文本的表示，进而增强预训练模型在一些知识图谱有关任务的效果。

4、CoLAKE: Contextualized Language and Knowledge Embedding

这篇论文来源于复旦和亚马逊，其主要关注于如何使用知识图谱以增强预训练模型的效果。

本文首先将上下文看作全连接图，并根据句子中的实体在KG上抽取子图，通过两个图中共现的实体将全连接图和KG子图融合起来；然后本文将该图转化为序列，使用Transformer进行预训练，并在训练时采用特殊的type embedding来表示实体、词语与其他子图信息。

5、EXPLOITING STRUCTURED KNOWLEDGE IN TEXT VIA GRAPH-GUIDED REPRESENTATION LEARNING

这篇论文来源于悉尼科技大学和微软，其主要关注于如何使用知识图谱增强预训练模型。

**四、中文自然语言处理相关资料**

<https://github.com/crownpku/Awesome-Chinese-NLP>

### Toolkits 综合NLP工具包

* [THULAC 中文词法分析工具包](http://thulac.thunlp.org/) by 清华 (C++/Java/Python)
* [NLPIR](https://github.com/NLPIR-team/NLPIR) by 中科院 (Java)
* [LTP 语言技术平台](https://github.com/HIT-SCIR/ltp) by 哈工大 (C++) [pylyp](https://github.com/HIT-SCIR/pyltp) LTP的python封装
* [FudanNLP](https://github.com/FudanNLP/fnlp) by 复旦 (Java)
* [BaiduLac](https://github.com/baidu/lac) by 百度 Baidu's open-source lexical analysis tool for Chinese, including word segmentation, part-of-speech tagging & named entity recognition.
* [HanLP](https://github.com/hankcs/HanLP) (Java)
* [FastNLP](https://github.com/fastnlp/fastNLP) (Python) 一款轻量级的 NLP 处理套件。
* [SnowNLP](https://github.com/isnowfy/snownlp) (Python) Python library for processing Chinese text
* [YaYaNLP](https://github.com/Tony-Wang/YaYaNLP) (Python) 纯python编写的中文自然语言处理包，取名于“牙牙学语”
* [小明NLP](https://github.com/SeanLee97/xmnlp) (Python) 轻量级中文自然语言处理工具
* [DeepNLP](https://github.com/rockingdingo/deepnlp) (Python) Deep Learning NLP Pipeline implemented on Tensorflow with pretrained Chinese models.
* [chinese\_nlp](https://github.com/taozhijiang/chinese_nlp) (C++ & Python) Chinese Natural Language Processing tools and examples
* [lightNLP](https://github.com/smilelight/lightNLP) (Python) 基于Pytorch和torchtext的自然语言处理深度学习框架
* [Chinese-Annotator](https://github.com/crownpku/Chinese-Annotator) (Python) Annotator for Chinese Text Corpus 中文文本标注工具
* [Poplar](https://github.com/synyi/poplar) (Typescript) A web-based annotation tool for natural language processing (NLP)
* [Jiagu](https://github.com/ownthink/Jiagu) (Python) Jiagu以BiLSTM等模型为基础，使用大规模语料训练而成。将提供中文分词、词性标注、命名实体识别、情感分析、知识图谱关系抽取、关键词抽取、文本摘要、新词发现等常用自然语言处理功能。
* [SmoothNLP](https://github.com/smoothnlp/SmoothNLP) (Python & Java) 专注于可解释的NLP技术
* [FoolNLTK](https://github.com/rockyzhengwu/FoolNLTK) (Python & Java) A Chinese Nature Language Toolkit

### Information Extraction 信息提取

* [MITIE](https://github.com/mit-nlp/MITIE) (C++) library and tools for information extraction
* [Duckling](https://github.com/facebookincubator/duckling) (Haskell) Language, engine, and tooling for expressing, testing, and evaluating composable language rules on input strings.
* [IEPY](https://github.com/machinalis/iepy) (Python) IEPY is an open source tool for Information Extraction focused on Relation Extraction.
* [Snorkel](https://github.com/HazyResearch/snorkel) A training data creation and management system focused on information extraction
* [Neural Relation Extraction implemented with LSTM in TensorFlow](https://github.com/thunlp/TensorFlow-NRE)
* [A neural network model for Chinese named entity recognition](https://github.com/zjy-ucas/ChineseNER)
* [bert-chinese-ner](https://github.com/ProHiryu/bert-chinese-ner) 使用预训练语言模型BERT做中文NER
* [Information-Extraction-Chinese](https://github.com/crownpku/Information-Extraction-Chinese) Chinese Named Entity Recognition with IDCNN/biLSTM+CRF, and Relation Extraction with biGRU+2ATT 中文实体识别与关系提取
* [Familia](https://github.com/baidu/Familia) 百度出品的 A Toolkit for Industrial Topic Modeling
* [Text Classification](https://github.com/brightmart/text_classification) All kinds of text classificaiton models and more with deep learning. 用知乎问答语聊作为测试数据。
* [ComplexEventExtraction](https://github.com/liuhuanyong/ComplexEventExtraction) 中文复合事件的概念与显式模式，包括条件事件、因果事件、顺承事件、反转事件等事件抽取，并形成事理图谱。
* [TextRank4ZH](https://github.com/letiantian/TextRank4ZH) 从中文文本中自动提取关键词和摘要

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 模型 | 语言模型 | 特征抽取 | 上下文表征 | 特点 | 链接 |
| 单向模型 自回归：标准LM | | ELMO | BiLM | BiLSTM | 单向 | 2个单向语言模型拼接 | flairNLP:https://github.com/flairNLP/flair |
| ULMFiT | LM | AWD-LSTM | 单向 | 引入逐层解冻,解决finetune中的灾难性问题 |  |
| SiATL | LM | LSTM | 单向 | 引入逐层解冻+辅助LM,解决finetune中的灾难性问题 |  |
| GPT1.0 | LM | Transformer | 单向 | 统一下游任务框架,验证transformer | flairNLP:https://github.com/flairNLP/flair |
| GPT2.0 | LM | Transformer | 单向 | 没有特定模型的精调流程，生成任务取得很好效果 | transformers:https://github.com/huggingface/transformers |
| BERT系列 自编码：MLM | BERT | BERT | MLM | Transformer | 双向 | MLM获取上下文相关的双向特征表示 | https://github.com/google-research/bert |
| 改进生成任务 | MASS | LM+MLM | Transformer | 单向/双向 | 有利于 Seq2Seq 类型的下游任务，比如 QA，总结和机器翻译 |  |
| UniLM | E-MLM | Transformer | 单向/双向 | 将对 mask 的预测扩展到三种任务：单向、双向和 Seq2Seq |  |
| 引入知识 | ERNIE1.0 | MLM(BPE) | Transformer | 双向 | 引入知识：3种[mask]策略（BPE）预测短语和实体 | <https://github.com/PaddlePaddle/ERNIE> |
| ERNIE(THU) | MLM+DEA | Transformer | 双向 | 引入知识：将实体向量与文本表示融合 | https://github.com/PaddlePaddle/ERNIE |
| 引入多任务 | MTDNN | MLM | Transformer | 双向 | 引入多任务学习：在下游阶段 |  |
| ERNIE2.0 | MLM+Multi-Task | Transformer | 双向 | 引入多任务学习：在预训练阶段，连续增量学习 | <https://github.com/PaddlePaddle/ERNIE> |
| 改进mask | SpanBERT | MLM+SPO | Transformer | 双向 | 不需要按照边界信息进行mask | <https://github.com/facebookresearch/SpanBERT> |
| 精细调参 | RoBERTa | E-MLM | Transformer | 双向 | 使用动态masking，精细调参，舍弃NSP | https://github.com/pytorch/fairseq/tree/master/examples/roberta |
|  | AlBert |  |  |  |  | <https://github.com/google-research/AlBert> |
| 广义自回归:PLM | | XLNet | PLM | Transformer-XL | 双向 | 排列语言模型+双注意力流+Transformer | https://github.com/zihangdai/xlnet |
| 其它未归类 | | T5 |  |  |  | 文本-文本 转换 | https://github.com/google-research/text-to-text-transfer-transformer |
| 中文预训练模型 | | Bert(for Chinese) | |  |  |  | https://github.com/ymcui/Chinese-BERT-wwm |
| RoBERTa(for Chinese) | |  |  |  | https://github.com/ymcui/Chinese-BERT-wwm |
| MacBERT(for Chinese) | |  |  |  | https://github.com/ymcui/MacBERT |
| XLNet(for Chinese) | |  |  |  | https://github.com/ymcui/Chinese-XLNet |
| ULMFiT(for Chinese) | |  |  |  | https://github.com/bigboNed3/chinese\_ulmfit |