医疗 NLP 领域（主要关注中文） 评测数据集 与 论文等相关资源。

- Chinese_medical_NLP
  - 中文评测数据集
    - 1. Yidu-S4K：医渡云结构化 4K 数据集
    - 2.瑞金医院糖尿病数据集
    - 3.Yidu-N7K：医渡云标准化 7K 数据集
    - 4.中文医学问答数据集
    - 5.平安医疗科技疾病问答迁移学习比赛
    - 6.天池新冠肺炎问句匹配比赛
    - 7.中文医患问答对话数据
    - 8.中文医学问答数据
  - 中文医学知识图谱
    - CMeKG
  - 英文数据集
    - PubMedQA: A Dataset for Biomedical Research Question Answering
  - 相关论文
    - 1.医疗领域预训练 embedding
    - 2.综述类文章
    - 3.电子病历相关文章
    - 4.医学关系抽取
    - 5.医学知识图谱
    - 6.辅助诊断
    - 7.ACL2020 医学领域相关论文列表
    - 8.医疗实体 Linking（标准化）
    - 9. AAAI2020 医学 NLP 相关论文列表
  - 中文医疗领域语料
    - 医学教材 培训考试
    - 哈工大《大词林》开放 75 万核心实体词及相关概念、关系列表（包含中药/医院/生物 类别）
  - 医学 embedding
    - 开源英文医学 embedding
  - 开源工具包

中文评测数据集

1. Yidu-S4K：医渡云结构化 4K 数据集

数据集描述：

> Yidu-S4K 数据集源自 CCKS 2019 评测任务一，即“面向中文电子病历的命名实体识别”的数据集，包括两个子任务： 1）医疗命名实体识别：由于国内没有公开可获得的面向中文电子病历医疗实体识别数据集，本年度保留了医疗命名实体识别任务，对 2017 年度数据集做了修订，并随任务一同发布。本子任务的数据集包括训练集和测试集。 2）医疗实体及属性抽取（跨院迁移）：在医疗实体识别的基础上，对预定义实体属性进行抽取。本任务为迁移学习任务，即在只提供目标场景少量标注数据的情况下，通过其他场景的标注数据及非标注数据进行目标场景的识别任务。本子任务的数据集包括训练集（非目标场景和目标场景的标注数据、各个场景的非标注数据）和测试集（目标场景的标注数据

数据集地址

度盘下载地址：https://pan.baidu.com/s/1QqYtqDwhc_S51F3SYMChBQ

提取码：flql

2.瑞金医院糖尿病数据集

数据集描述：

> 数据集来自天池大赛。此数据集旨在通过糖尿病相关的教科书、研究论文来做糖尿病文献挖掘并构建糖尿病知识图谱。参赛选手需要设计高准确率，高效的算法来挑战这一科学难题。第一赛季课题为“基于糖尿病临床指南和研究论文的实体标注构建”，第二赛季课题为“基于糖尿病临床指南和研究论文的实体间关系构建”。

官方提供的数据只包含训练集，真正用于最终排名的测试集没有给出。

数据集地址

度盘下载地址：https://pan.baidu.com/s/1CWKblBNBqR-vs2h0xiXSdQ

提取码：0c54

3.Yidu-N7K：医渡云标准化 7K 数据集

数据集描述：

> Yidu-N4K 数据集源自 CHIP 2019 评测任务一，即“临床术语标准化任务”的数据集。 临床术语标准化任务是医学统计中不可或缺的一项任务。临床上，关于同一种诊断、手术、药品、检查、化验、症状等往往会有成百上千种不同的写法。标准化（归一）要解决的问题就是为临床上各种不同说法找到对应的标准说法。有了术语标准化的基础，研究人员才可对电子病历进行后续的统计分析。

本质上，临床术语标准化任务也是语义相似度匹配任务的一种。但是由于原词表述方式过于多样，单一的匹配模型很难获得很好的效果。

数据集地址

4.中文医学问答数据集

数据集描述：

中文医药方面的问答数据集，超过 10 万条。

数据说明:

questions.csv：所有的问题及其内容。answers.csv ：所有问题的答案。 train_candidates.txt， dev_candidates.txt，test_candidates.txt ：将上述两个文件进行了拆分。

数据集地址

数据集 github 地址

5.平安医疗科技疾病问答迁移学习比赛

数据集描述：

本次比赛是 chip2019 中的评测任务二，由平安医疗科技主办。chip2019 会议详情见链接：http://cips-chip.org.cn/evaluation 迁移学习是自然语言处理中的重要一环，其主要目的是通过从已学习的相关任务中转移知识来改进新任务的学习效果，从而提高模型的泛化能力。 本次评测任务的主要目标是针对中文的疾病问答数据，进行病种间的迁移学习。具体而言，给定来自 5 个不同病种的问句对，要求判定两个句子语义是否相同或者相近。所有语料来自互联网上患者真实的问题，并经过了筛选和人工的意图匹配标注。

数据集地址(需注册)

6.天池新冠肺炎问句匹配比赛

数据集描述：

本次大赛数据包括：脱敏之后的医疗问题数据对和标注数据。医疗问题涉及“肺炎”、“支原体肺炎”、“支气管炎”、“上呼吸道感染”、“肺结核”、“哮喘”、“胸膜炎”、“肺气肿”、“感冒”、“咳血”等 10 个病种。 数据共包含 train.csv、dev.csv、test.csv 三个文件，其中给参赛选手的文件包含训练集 train.csv 和验证集 dev.csv，测试集 test.csv 对参赛选手不可见。 每一条数据由 Category，Query1，Query2，Label 构成，分别表示问题类别、问句 1、问句 2、标签。Label 表示问句之间的语义是否相同，若相同，标为 1，若不相同，标为 0。其中，训练集 Label 已知，验证集和测试集 Label 未知。 示例 类别：肺炎 问句 1：肺部发炎是什么原因引起的？ 问句 2：肺部发炎是什么引起的 标签:1 类别：肺炎 问句 1：肺部发炎是什么原因引起的？ 问句 2：肺部炎症有什么症状 标签:0

数据集地址(需注册)

线上第四名解决方案及代码

线上第一名解决方案及代码

7.中文医患问答对话数据

数据说明: 来自某在线求医产品的中文医患对话数据。

原始描述:The MedDialog dataset contains conversations (in Chinese) between doctors and patients. It has 1.1 million dialogues and 4 million utterances. The data is continuously growing and more dialogues will be added. The raw dialogues are from haodf.com. All copyrights of the data belong to haodf.com.

项目地址

度盘下载地址: https://pan.baidu.com/s/1ZwzNgvAAMQk4klerTspsoA

提取码: lbo4

8.中文医学问答数据

数据说明: 包含六个科室的医学问答数据，来源不明。

项目地址

中文医学知识图谱

CMeKG

地址

简介：CMeKG（Chinese Medical Knowledge Graph）是利用自然语言处理与文本挖掘技术，基于大规模医学文本数据，以人机结合的方式研发的中文医学知识图谱。CMeKG 的构建参考了 ICD、ATC、SNOMED、MeSH 等权威的国际医学标准以及规模庞大、多源异构的临床指南、行业标准、诊疗规范与医学百科等医学文本信息。CMeKG 1.0 包括：6310 种疾病、19853 种药物（西药、中成药、中草药）、1237 种诊疗技术及设备的结构化知识描述，涵盖疾病的临床症状、发病部位、药物治疗、手术治疗、鉴别诊断、影像学检查、高危因素、传播途径、多发群体、就诊科室等以及药物的成分、适应症、用法用量、有效期、禁忌证等 30 余种常见关系类型，CMeKG 描述的概念关系实例及属性三元组达 100 余万。

英文数据集

PubMedQA: A Dataset for Biomedical Research Question Answering

数据集描述： 基于 Pubmed 提取的医学问答数据集。PubMedQA has 1k expert-annotated, 61.2k unlabeled and 211.3k artificially gen- erated QA instances.

论文地址

相关论文

1.医疗领域预训练 embedding

注：目前没有收集到中文医疗领域的开源预训练模型，以下列出英文论文供参考。

Bio-bert

论文题目：BioBERT: a pre-trained biomedical language representation model for biomedical text mining

论文地址

项目地址

论文概要：以通用领域预训练 bert 为初始权重，基于 Pubmed 上大量医疗领域英文论文训练。在多个医疗相关下游任务中超越 SOTA 模型的表现。

论文摘要：

Motivation: Biomedical text mining is becoming increasingly important as the number of biomedical documents rapidly grows. With the progress in natural language processing (NLP), extracting valuable information from bio- medical literature has gained popularity among researchers, and deep learning has boosted the development of ef- fective biomedical text mining models. However, directly applying the advancements in NLP to biomedical text min- ing often yields unsatisfactory results due to a word distribution shift from general domain corpora to biomedical corpora. In this article, we investigate how the recently introduced pre-trained language model BERT can be adapted for biomedical corpora. **Results**: We introduce BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. With almost the same architecture across tasks, BioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text mining tasks when pre-trained on biomedical corpora. While BERT obtains performance comparable to that of previous state-of-the-art models, BioBERT significantly outperforms them on the following three representative biomedical text mining tasks: biomedical named entity recognition (0.62% F1 score improvement), biomedical relation extraction (2.80% F1 score improvement) and biomedical question answering (12.24% MRR improvement). Our analysis results show that pre-training BERT on biomedical corpora helps it to understand complex biomedical texts. **Availability and implementation**: We make the pre-trained weights of BioBERT freely available at https://github.com/naver/biobert-pretrained, and the source code for fine-tuning BioBERT available at https://github.com/dmis-lab/biobert.

sci-bert

论文题目：SCIBERT: A Pretrained Language Model for Scientific Text

论文地址

项目地址

论文概要：AllenAI 团队出品,基于 Semantic Scholar 上 110 万+ 文章训练的 科学领域 bert.

论文摘要：Obtaining large-scale annotated data for NLP tasks in the scientific domain is challeng- ing and expensive. We release SCIBERT, a pretrained language model based on BERT (Devlin et al., 2019) to address the lack of high-quality, large-scale labeled scientific data. SCIBERT leverages unsupervised pretraining on a large multi-domain corpus of scientific publications to improve perfor- mance on downstream scientific NLP tasks. We evaluate on a suite of tasks including sequence tagging, sentence classification and dependency parsing, with datasets from a variety of scientific domains. We demon- strate statistically significant improvements over BERT and achieve new state-of-the- art results on several of these tasks. The code and pretrained models are available at https://github.com/allenai/scibert/.

clinical-bert

论文题目：Publicly Available Clinical BERT Embeddings

论文地址

项目地址

论文概要：出自 NAACL Clinical NLP Workshop 2019.基于 MIMIC-III 数据库中的 200 万份医疗记录训练的临床领域 bert.

论文摘要：Contextual word embedding models such as ELMo and BERT have dramatically improved performance for many natural language processing (NLP) tasks in recent months. However, these models have been minimally explored on specialty corpora, such as clinical text; moreover, in the clinical domain, no publicly-available pre-trained BERT models yet exist. In this work, we address this need by exploring and releasing BERT models for clinical text: one for generic clinical text and another for discharge summaries specifically. We demonstrate that using a domain-specific model yields performance improvements on 3/5 clinical NLP tasks, establishing a new state-of-the-art on the MedNLI dataset. We find that these domain-specific models are not as performant on 2 clinical de-identification tasks, and argue that this is a natural consequence of the differences between de-identified source text and synthetically non de-identified task text.

clinical-bert(另一团队的版本)

论文题目：ClinicalBert: Modeling Clinical Notes and Predicting Hospital Readmission

论文地址

项目地址

论文概要：同样基于 MIMIC-III 数据库,但只随机选取了 10 万份医疗记录训练的临床领域 bert.

论文摘要：Clinical notes contain information about patients that goes beyond structured data like lab values and medications. However, clinical notes have been underused relative to structured data, because notes are high-dimensional and sparse. This work develops and evaluates representations of clinical notes using bidirectional transformers (ClinicalBert). Clini- calBert uncovers high-quality relationships between medical concepts as judged by hu- mans. ClinicalBert outperforms baselines on 30-day hospital readmission prediction using both discharge summaries and the first few days of notes in the intensive care unit. Code and model parameters are available.

BEHRT

论文题目：BEHRT: TRANSFORMER FOR ELECTRONIC HEALTH RECORDS

论文地址

项目地址: 暂未开源

论文概要：这篇论文中 embedding 是基于医学实体训练，而不是基于单词。

论文摘要：Today, despite decades of developments in medicine and the growing interest in precision healthcare, vast majority of diagnoses happen once patients begin to show noticeable signs of illness. Early indication and detection of diseases, however, can provide patients and carers with the chance of early intervention, better disease management, and efficient allocation of healthcare resources. The latest developments in machine learning (more specifically, deep learning) provides a great opportunity to address this unmet need. In this study, we introduce BEHRT: A deep neural sequence transduction model for EHR (electronic health records), capable of multitask prediction and disease trajectory mapping. When trained and evaluated on the data from nearly 1.6 million individuals, BEHRT shows a striking absolute improvement of 8.0-10.8%, in terms of Average Precision Score, compared to the existing state-of-the-art deep EHR models (in terms of average precision, when predicting for the onset of 301 conditions). In addition to its superior prediction power, BEHRT provides a personalised view of disease trajectories through its attention mechanism; its flexible architecture enables it to incorporate multiple heterogeneous concepts (e.g., diagnosis, medication, measurements, and more) to improve the accuracy of its predictions; and its (pre-)training results in disease and patient representations that can help us get a step closer to interpretable predictions.

2.综述类文章

nature medicine 发表的综述

论文题目：A guide to deep learning in healthcare

论文地址

论文概要：发表于 nature medicine，包含医学领域下 CV,NLP,强化学习等方面的应用综述。

论文摘要：Here we present deep-learning techniques for healthcare, centering our discussion on deep learning in computer vision, natural language processing, reinforcement learning, and generalized methods. We describe how these computational techniques can impact a few key areas of medicine and explore how to build end-to-end systems. Our discussion of computer vision focuses largely on medical imaging, and we describe the application of natural language processing to domains such as electronic health record data. Similarly, reinforcement learning is discussed in the context of robotic-assisted surgery, and generalized deep- learning methods for genomics are reviewed.

3.电子病历相关文章

Transfer Learning from Medical Literature for Section Prediction in Electronic Health Records

论文地址

论文概要：发表于 EMNLP2019。基于少量 in-domain 数据和大量 out-of-domain 数据进行 EHR 相关的迁移学习。

论文摘要：sections such as Assessment and Plan, So- cial History, and Medications. These sec- tions help physicians find information easily and can be used by an information retrieval system to return specific information sought by a user. However, it is common that the exact format of sections in a particular EHR does not adhere to known patterns. There- fore, being able to predict sections and headers in EHRs automatically is beneficial to physi- cians. Prior approaches in EHR section pre- diction have only used text data from EHRs and have required significant manual annota- tion. We propose using sections from med- ical literature (e.g., textbooks, journals, web content) that contain content similar to that found in EHR sections. Our approach uses data from a different kind of source where la- bels are provided without the need of a time- consuming annotation effort. We use this data to train two models: an RNN and a BERT- based model. We apply the learned models along with source data via transfer learning to predict sections in EHRs. Our results show that medical literature can provide helpful su- pervision signal for this classification task.

4.医学关系抽取

Leveraging Dependency Forest for Neural Medical Relation Extraction

论文地址

论文概要：发表于 EMNLP 2019. 基于 dependency forest 方法，提升对医学语句中依存关系的召回率，同时引进了一部分噪声，基于图循环网络进行特征提取，提供了在医疗关系抽取中使用依存关系，同时减少误差传递的一种思路。

论文摘要：Medical relation extraction discovers relations between entity mentions in text, such as research articles. For this task, dependency syntax has been recognized as a crucial source of features. Yet in the medical domain, 1-best parse trees suffer from relatively low accuracies, diminishing their usefulness. We investigate a method to alleviate this problem by utilizing dependency forests. Forests contain more than one possible decisions and therefore have higher recall but more noise compared with 1-best outputs. A graph neural network is used to represent the forests, automatically distinguishing the useful syntactic information from parsing noise. Results on two benchmarks show that our method outperforms the standard tree-based methods, giving the state-of-the-art results in the literature.

5.医学知识图谱

Learning a Health Knowledge Graph from Electronic Medical Records

论文地址

论文概要：发表于 nature scientificreports（2017）. 基于 27 万余份电子病历构建的疾病-症状知识图谱。

论文摘要：Demand for clinical decision support systems in medicine and self-diagnostic symptom checkers has substantially increased in recent years. Existing platforms rely on knowledge bases manually compiled through a labor-intensive process or automatically derived using simple pairwise statistics. This study explored an automated process to learn high quality knowledge bases linking diseases and symptoms directly from electronic medical records. Medical concepts were extracted from 273,174 de-identified patient records and maximum likelihood estimation of three probabilistic models was used to automatically construct knowledge graphs: logistic regression, naive Bayes classifier and a Bayesian network using noisy OR gates. A graph of disease-symptom relationships was elicited from the learned parameters and the constructed knowledge graphs were evaluated and validated, with permission, against Google's manually-constructed knowledge graph and against expert physician opinions. Our study shows that direct and automated construction of high quality health knowledge graphs from medical records using rudimentary concept extraction is feasible. The noisy OR model produces a high quality knowledge graph reaching precision of 0.85 for a recall of 0.6 in the clinical evaluation. Noisy OR significantly outperforms all tested models across evaluation frameworks ($p < 0.01$).

6.辅助诊断

Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence

论文地址

论文概要：该文章由广州市妇女儿童医疗中心与依图医疗等企业和科研机构共同完成，基于机器学习的自然语言处理（NLP）技术实现不输人类医生的强大诊断能力，并具备多场景的应用能力。据介绍，这是全球首次在顶级医学杂志发表有关自然语言处理（NLP）技术基于电子健康记录（EHR）做临床智能诊断的研究成果，也是利用人工智能技术诊断儿科疾病的重磅科研成果。

论文摘要：Artificial intelligence (AI)-based methods have emerged as powerful tools to transform medical care. Although machine learning classifiers (MLCs) have already demonstrated strong performance in image-based diagnoses, analysis of diverse and massive electronic health record (EHR) data remains challenging. Here, we show that MLCs can query EHRs in a manner similar to the hypothetico-deductive reasoning used by physicians and unearth associations that previous statistical methods have not found. Our model applies an automated natural language processing system using deep learning techniques to extract clinically relevant information from EHRs. In total, 101.6 million data points from 1,362,559 pediatric patient visits presenting to a major referral center were analyzed to train and validate the framework. Our model demonstrates high diagnostic accuracy across multiple organ systems and is comparable to experienced pediatricians in diagnosing common childhood diseases. Our study provides a proof of concept for implementing an AI-based system as a means to aid physicians in tackling large amounts of data, augmenting diagnostic evaluations, and to provide clinical decision support in cases of diagnostic uncertainty or complexity. Although this impact may be most evident in areas where healthcare providers are in relative shortage, the benefits of such an AI system are likely to be universal.

7.ACL2020 医学领域相关论文列表

A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization

论文地址

Biomedical Entity Representations with Synonym Marginalization

论文地址

Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain

论文地址

MIE: A Medical Information Extractor towards Medical Dialogues

论文地址

Rationalizing Medical Relation Prediction from Corpus-level Statistics

论文地址

8.医疗实体 Linking（标准化）

Medical Entity Linking using Triplet Network

论文地址

论文概要： 发表于 ACL2019,论文内容为疾病实体 Linking 研究。使用三元组数据，（mention，正例，负例），目标使 distance(mention,负例)-distance(mention,正例)>alpha（人脸识别的经典方案）,具体损失函数参看论文。论文主要包括两部分内容 1）候选数据集生成,对给定 mention，与标准疾病集合数据（标准词及同义词）计算余弦相似度及 Jaccard overlap 分数,取 topK 作为候选样例。 2）网络结构基于 Triplet Network。详见论文。

A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization

论文地址

论文概要: 发表于 ACL2020。基于 list-wise 排序学习方法。主要分为两部分：后续数据集生成 和 基于 BERT 的 list-wise 排序。较新颖的思路：1）在样本生成过程中，对标准词进行了基于同义词的扩展。2）在 loss 中引入了语义类型正则化。详见论文。

9. AAAI2020 医学 NLP 相关论文列表

On the Generation of Medical Question-Answer Pairs

论文地址

LATTE: Latent Type Modeling for Biomedical Entity Linking

论文地址

Learning Conceptual-Contextual Embeddings for Medical Text

论文地址

Understanding Medical Conversations with Scattered Keyword Attention and Weak Supervision from Responses

论文地址

Simultaneously Linking Entities and Extracting Relations from Biomedical Text without Mention-level Supervision

论文地址

Can Embeddings Adequately Represent Medical Terminology? New Large-Scale Medical Term Similarity Datasets Have the Answer!

论文地址

中文医疗领域语料

医学教材+培训考试

说明:由于版权原因，现在无法提供度盘下载链接了，请大家前往原豆瓣链接下载吧。

语料说明：根据此豆瓣链接整理。

数据预览：

哈工大《大词林》开放 75 万核心实体词及相关概念、关系列表（包含中药/医院/生物 类别）

语料说明:哈工大开源了《大词林》中的 75 万的核心实体词，以及这些核心实体词对应的细粒度概念词（共 1.8 万概念词，300 万实体-概念元组），还有相关的关系三元组（共 300 万）。这 75 万核心实体列表涵盖了常见的人名、地名、物品名等术语。概念词列表则包含了细粒度的实体概念信息。借助于细粒度的上位概念层次结构和丰富的实体间关系，本次开源的数据能够为人机对话、智能推荐、等应用技术提供数据支持。

语料官方下载地址

说明: 通过网上查询，这部分资源应该是被一些公司付费使用了，可能有版权问题，所以现在下载链接都失效了。后续如果再有开源的信息再进行更新。

医学 embedding

开源英文医学 embedding

项目说明：发表于 AMIA 2016. 开源医学相关概念 embedding.

项目地址

开源工具包

分词工具

PKUSEG

项目地址

项目说明： 北京大学推出的多领域中文分词工具，支持选择医学领域。

工业级产品解决方案

灵医智慧

左手医生

blog 分享

医疗领域构建自然语言处理系统的经验教训