计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190747
57(11): 2467-2477, 2020

融合常用语的大规模疾病术语图谱构建

张晨童[1] 张佳影[1] 张知行[1] 阮 彤[1] 何 萍[2] 葛小玲[3]
[1](华东理工大学 上海 200237)
[2](上海申康医院发展中心 上海 200041)
[3](复旦大学附属儿科医院 上海 201108)
(chentong_zhang@163.com)

Construction of Large-Scale Disease Terminology Graph with Common Terms

Zhang Chentong[1], Zhang Jiaying[1], Zhang Zhixing[1], Ruan Tong[1], He Ping[2], and Ge Xiaoling[3]
[1](*East China University of Science and Technology*, *Shanghai* 200237)
[2](*Shanghai Hospital Development Center*, *Shanghai* 200041)
[3](*Children's Hospital of Fudan University*, *Shanghai* 201108)

Abstract The National Health Planning Commission requires medical institutions to use the ICD (international classification of diseases) codes. However, due to the large amount of common terms in clinical disease descriptions, the direct matching rate between clinical diagnostic names in electronic medical records and ICD codes is low. Based on the real diagnostic data on the regional healthcare platform, this paper constructs a disease terminology graph fusing common terms. Specifically, this paper proposes a relationship recognition algorithm based on data enhancement which combines the rule algorithm based on the disease components and the pre-training BERT (bidirectional encoder representation from transformers) model. The proposed algorithm identifies synonymy and hypernymy between over 50 000 common terms and diseases in ICD10(international classification of diseases 10th revision, Chinese version), then further fuses the hierarchical structure of ICD11 (international classification of diseases 11th revision, Chinese version). Moreover, this paper also proposes a task allocation algorithm based on the disease-department association graph to perform manual verification. Finally, a large-scale disease terminology graph including 1 460 synonyms and 46 508 hypernymy can be formed by 94 478 disease entities. The evaluation experiments show that the coverage of clinical diagnostic data based on disease terminology graph is 75.31% higher than direct mapping based on ICD10. In addition, compared with manual coding by doctors, the automatic coding using disease terminology graph can shorten 59.75% of the encoding time, and the accuracy rate is 85%.

Key words common terms; disease terminology graph; ICD(international classification of diseases); relationship recognition; verification

摘 要 国家卫计委要求医疗机构使用国际疾病分类(international classification of diseases, ICD)编码,然而由于临床疾病描述存在大量的常用词,导致电子病历中录入的诊断名称与ICD编码直接映射匹配率低.基于区域健康平台上的真实诊断数据,构建了融合常用语的疾病术语图谱.具体来说,在基于

收稿日期:2019-10-21;**修回日期**:2020-02-24
基金项目:国家自然科学基金项目(61772201);国家重点研发计划项目(2018YFC0910500)
This work was supported by the National Natural Science Foundation of China (61772201) and the National Key Research and Development Program of China (2018YFC0910500).
通信作者:阮彤(ruantong@ecust.edu.cn)

疾病构成成分的规则算法基础上,提出了基于数据增强的 BERT(bidirectional encoder representation from transformers)上下位关系识别算法,将 5 万多个诊断常用语和 ICD10(international classification of diseases 10th revision,Chinese version)中的疾病进行同义关系和上下位关系识别,进一步融合了 ICD11(international classification of diseases 11th revision,Chinese version)的层次结构,此外,还提出了基于疾病-科室关联图谱的任务分配方法以进行人工校验,最终 94 478 个疾病实体形成了包含 1 460 条同义关系、46 508 条上下位关系的大规模疾病术语图谱.评估实验表明,基于疾病术语图谱,对临床诊断数据的覆盖率比基于 ICD10 的直接映射编码的覆盖率提升了 75.31%,另外,利用疾病术语图谱自动进行编码疾病相比于医生人工编码会缩短约 59.75%的编码时间,且正确率达到 85%.

关键词 常用语;疾病术语图谱;国际疾病分类;关系识别;校验

中图法分类号 TP391

随着医疗领域信息化程度的不断提升,欧美国家的医疗研究机构已经建立起一系列医学术语库,如临床医疗术语集(systematized nomenclature of medicine-clinical terms, SNOMED-CT)①、一体化医学语言系统(unified medical language system, UMLS)②、ICD-10(international classification of diseases 10th revision)③、ICD-11(international classification of diseases 11th revision)④等.其中,中华人民共和国国家卫生健康委员会明确要求各医疗机构在病案书写中统一使用 ICD-10 中文版(简称 ICD10),这极大地推进了医疗服务规范化、标准化管理.

然而,ICD10 在实际应用于临床数据时,能够直接建立映射的比例不足 20%.其主要问题有 2 点:第一,疾病名称描述的多样性.例如,"尿路感染"是临床诊断中的常用语,但在 ICD10 中并未收录该词.该词是"泌尿道感染"的同义词,后者在 ICD10 中对应的编码为 N39.0.第二,疾病常用语的粒度更细.例如,"糖尿病伴有眼部改变"在 ICD10 中无法找到与之匹配的同义词,仅能找到它的上位词"糖尿病",糖尿病在 ICD10 中对应的编码为 E14.900.因此,以 ICD10 为标准,构建一个融合常用语的疾病术语图谱,将常用语作为同义词或下位词融入 ICD10,可以有效建立疾病名称与 ICD10 的映射关系,将方便医生查找疾病名称或是机器进行 ICD 自动编码.然而常用语的融合需要大量医学知识,而人工建立映射耗时耗力,机器自动映射准确率比较低.另外,ICD10 的分类体系延续了传统的列表式结构,过于平面,并不方便浏览与查找.

针对上述问题与难点,本文提出了一种融合常用语的大规模疾病术语图谱构建方案.具体来说,本文筛选出上海市区域医疗健康平台(其包含市内 38 家三级医院的临床诊疗信息)的疾病数据中的常用语,将常用语与 ICD10 进行融合.此外,为方便医生查找,将 ICD10 的类目层与 ICD-11 中文版(简称 ICD11)的层次结构进行进一步融合,形成融合常用语的大规模疾病术语图谱.本方案的贡献点有 3 个方面:

1) 疾病术语图谱的构建将机器与人工的优点结合起来.首先分析疾病词构成成分,利用基于疾病构成成分的规则算法识别出疾病间的同义关系,并通过基于数据增强的 BERT(bidirectional encoder representation from transformers)上下位关系识别算法找出疾病间的上下位关系,进而利用 ICD 体系本身特点,按照疾病类型,基于专科分组校验疾病数据.本文还设计了基于疾病-科室关联图谱的任务分配方法,方便校对人员对医疗数据进行校验,以保证疾病医学实体关系的准确度.

2) 面向临床诊断数据构建出了融合常用语的大规模疾病术语图谱,图谱可以表示出医疗术语间的上下位关系和同义关系,将常用语与标准术语融合起来.最终找出 1 460 条同义关系、46 508 条上下位关系.

3) 本文构建的疾病术语图谱,在维护现有标准体系的同时,兼顾了临床使用的方便性.本文从编码覆盖率、编码效率和编码正确率 3 方面对疾病术语图谱进行了评估,利用本文构建的疾病术语图谱相比于 ICD10 体系能够平均多覆盖 75.31%的临床诊断数据,并且利用疾病术语图谱辅助编码相比于人工编码,能够缩短约 59.75%的时间,且正确率达到 85%.

① http://www.snomed.org/snomed-ct/

② https://www.nlm.nih.gov.research.umls

③ https://icd.who.int/browse10/2016/en

④ https://icd.who.int/browse11/l-m/en

1 相关工作

国内外关于术语体系构建方面有很丰富的研究.国外构建了大量的生物医疗分类体系,除了UMLS[1],SNOMED-CT 等通用的分类系统,还有面向药物的命名系统 RxNorm①、针对检验的编码系统 LOINC②和被广泛应用的国际疾病分类系统等细分的系统.而国内在医学术语体系上不断和国际接轨,如 ICD10.

早期术语体系的构建采用纯手工的方式,如面向语义的英语词典 WordNet[2]和常识知识图谱 CYC[3],其中 CYC 由 50 万实体、700 万条断言构成.近年来,采用自动方法构建术语体系得到广泛应用,其构建过程涉及自动分类归纳的问题,即能够有效地扩充整个知识结构,有大量的工作研究了基于语言模型匹配的方法,用来解决术语与其上位词之间关系的自动分类归纳问题.如 Hearst[4]描述了一种从无限制文本中自动获取下位词的方法,确定了一组易于识别的词汇-句法模式.Navigli 等人[5]提出了一种基于图形的方法,旨在从域语料库和 Web 开始自动学习词汇分类法.实验表明,无论是在构建全新的分类法时还是在重构 WordNet 子层次结构时,都能获得高质量的结果.Snow 等人[6]提出了一种从文本中自动学习上下位(*is-a*)关系的新算法来解决自动构建和扩展语义分类法(如 WordNet)的问题.Yang 等人[7]提出了一种新的基于度量的框架,用于自动分类归纳的任务.近年来,使用基于字嵌入的方法来识别关系以重建分类法的做法也十分普及[8-11].

将常用术语等新信息加入到现存分类法中,主要集中在增强 WordNet 分类标准上.Toral 等人[12]丰富了 WordNet 的 310 742 个命名实体和 381 043 个"关系实例".Fellbaum 等人[13]创建了 Medical-WordNet,其不仅是对原始 WordNet 中医学术语的词汇扩展,而是提出了一种新型的存储库.Vedula 等人[14]研究了知识结构扩充问题,即对于大量出现的新的概念,怎么将其添加到已有的知识结构中.这一问题存在双重挑战,如何检测未知的实体或概念,以及新的概念怎样插入到已有的知识结构中而不破坏新创建的关系的语义完整性.他们提出了 ETF 的框架,用来自新闻和研究出版物等资源的新概念来丰富大规模的通用分类法,将新概念链接到现有概念上,获得潜在的父子关系.

然而,单纯采用人工构建的方式需要耗费大量的人力物力,仅使用自动构建的方式又不能保证机器的正确率.因此本文采用了人工和自动构建相结合的方法.

2 疾病术语图谱构建

2.1 问题定义

本文参考并扩展了 ICD10 及 ICD11 的分类层级体系,将疾病医学实体关系定义为:

定义 1. 不同疾病医学实体之间的映射关系 $R(E_i,E_j)$.其中 E_i,E_j 为疾病医学实体,R 为映射关系.映射关系包括 2 类:

1) $is\_hypernym$ 关系.$is\_hypernym(E_i,E_j)$ 表示实体 E_i 和 E_j 之间的上下位关系.特别地,$is\_hypernym$ 关系具有反函数性:$is\_hypernym(E_i,E_j)\Leftrightarrow is\_hyponymy(E_j,E_i)$,即 E_i 是 E_j 的上位词,等价于 E_j 是 E_i 的下位词.为了方便,如若不做特别说明,本文此后的上下位关系专指上位关系.

2) is_same 关系.$is\_same(E_i,E_j)$ 表示实体 E_i 和 E_j 之间的同义关系.同义关系包括 2 部分:一是医学上的同义关系,类似于"胰岛素依赖型糖尿病"和"1 型糖尿病"属于同义关系.二是因医生书写习惯不同导致的同义关系,类似于"1 型糖尿病"与"糖尿病(1 型)"属于同义关系.

本文的主要任务,是将常用语根据疾病医学实体关系链接上 ICD10,并且将 ICD10 中类目层与 ICD11 的层次结构进行融合,以构建出融合常用语的大规模疾病术语图谱.其中,常用语定义为区域平台上临床诊断疾病数据中出现频次大于 5 次的疾病名称.

2.2 整体框架

本文的整体框架如图 1 所示.首先 ICD10 融合常用语,然后再添加 ICD11 层次结构信息,最终形成融合常用语的疾病术语图谱.

图 1 左侧展示了疾病术语图谱的基本框架,首先将常用语与 ICD10 中标准疾病名称融合,融合过程即判断疾病对〈ICD10 中标准疾病术语,常用语〉是否具有上下位关系或同义关系,根据疾病对的疾病

① https://www.nlm.nih.gov/research/umls/rxnorm/docs/prescribe.html

② https://loinc.org

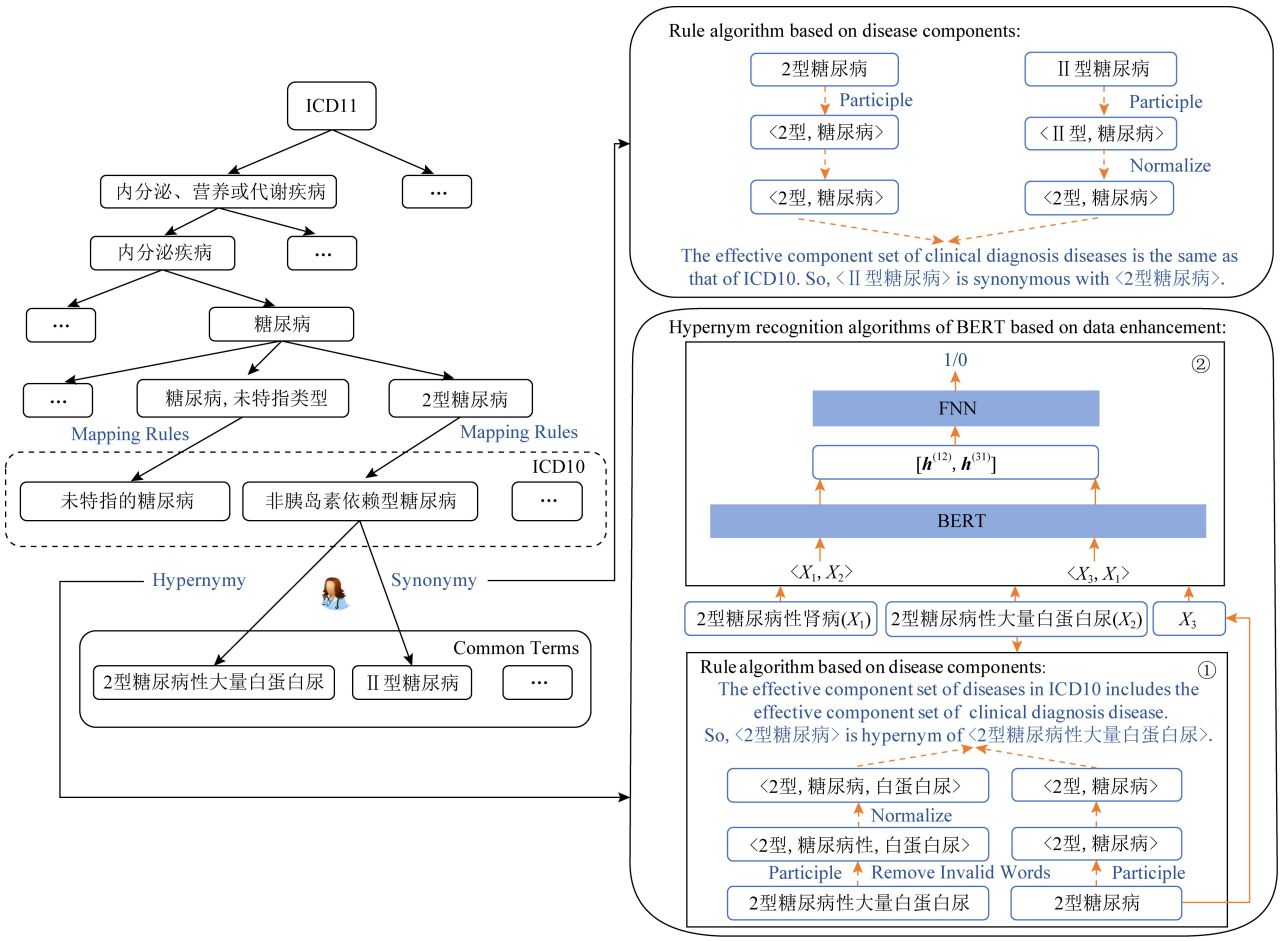

Fig. 1 Overall framework of large-scale diseases terminology graph by integrating common terms

图 1 融合常用语的大规模疾病术语图谱整体框架

医学实体关系，将常用语链到 ICD10 各层上，实现对常用语的归类.图 1 右侧分别显示了利用疾病构成成分的规则算法识别疾病对是否具有同义关系以及在基于疾病构成成分的规则算法基础上，结合 BERT 识别疾病对是否具有上下位关系.其次，根据映射规则将 ICD10 中类目层链接到 ICD11 的层级结构上.最后，为了保证融合结果的正确性，引入了基于疾病-科室关联图谱的任务分配方法，方便校验人员对疾病术语图谱包含的疾病医学实体关系进行修正处理.

2.3 ICD10 融合常用语

对于疾病术语之间的关系识别任务，本文定义为同义和上下位关系识别，而其重点在于上下位关系的识别.Wang 等人[15]提出基于规则的上下位识别算法，由知识驱动，通过预先构建包含大量细粒度临床实体的词典以及实体间上下位关系的集合，用以关系判断.基于规则的方法能够高质量地识别出上下位关系，然而受限于词典的规模，其召回率很低.因此，本文在使用预训练模型的基础上，结合规则提供的参考结果予以辅助信息，提出基于数据增强的 BERT 上下位关系识别算法.

给定一个疾病对$\langle X_1, X_2\rangle$，其中，X_1 是 ICD10 中的标准疾病术语，X_2 是常用语.首先将 X_2 送入基于疾病构成成分的规则算法，得到 X_2 在 ICD10 全集中的最优匹配词 X_3，构造最优匹配对$\langle X_2, X_3\rangle$.接着，由$\langle X_1, X_2\rangle$和$\langle X_2, X_3\rangle$，生成参考对$\langle X_3, X_1\rangle$.然后，疾病对$\langle X_1, X_2\rangle$和参考对$\langle X_3, X_1\rangle$分别经过 BERT[16]，得到词对中 2 个元素的相关性表示 $\boldsymbol{h}^{(12)}$ 和 $\boldsymbol{h}^{(31)}$.最后，拼接 $\boldsymbol{h}^{(12)}$ 和 $\boldsymbol{h}^{(31)}$，使用前馈神经网络(feedforward neural network, FNN)进行上下位关系预测.

本文将常用语与所有 ICD10 中标准疾病术语组成疾病对，利用基于数据增强的 BERT 上下位关系识别模型对疾病对进行上下位关系的预测，并对所有的预测结果根据模型预测为上下位关系的概率进行排序，将最高概率的$\langle X_1, X_2\rangle$作为最终输出结果.

2.3.1 参考对构造

构造参考对$\langle X_3, X_1\rangle$的目标在于,通过判断基于疾病构成成分的规则算法所预测的参考结果 X_3 和 X_1 之间的相关性,为疾病对$\langle X_1, X_2\rangle$的上下位关系识别提供辅助信息.因此,本文定义疾病构成成分来获得相应的词典,并利用基于疾病构成成分的规则算法得到 X_3.

1) 疾病构成成分定义

基于对 ICD 及区域医疗平台包含的 38 家医院的临床诊断数据的分析,本文将疾病词归纳为由原子疾病词(atomic disease words)、病因词(causal words)、病理词(pathological words)、部位词(part words)、临床表现词(clinical expression words)5 大成分构成,表 1 给出了具体含义.

Table 1 Examples of Disease Components

表 1 疾病构成成分举例

Disease Components	Meaning
Atomic Disease Words	A part of a disease name, but not divided into finer-grained words. Such as "糖尿病".
Causal Words	Including the cause of disease and conditions. The cause of disease refers to those factors that can cause the disease and give the disease specificity. For example,"遗传性"is the causal word of "遗传性聋".
Pathological Words	Modifying words such as severity, nature and period of onset. For example,"妊娠期"is the pathological word of "妊娠期高血压".
Part Words	Indicating the location of disease in disease name. For example,"胃"is the part word of"胃溃疡".
Clinical Expression Words	A series of abnormal changes in a patient's body after he has a certain disease. Such as"发热".

2) 基于疾病构成成分的规则算法

给定 ICD10 的疾病名称集合 $D=\{D_1, D_2, \cdots, D_n\}$,其中 n 为疾病名称的总个数.基于疾病构成成分的规则算法先基于疾病构成成分的双向最大匹配算法分别对 D_i 和 X_2 进行分词,剔除其中的无效词,无效词具体包括连接词、副词、程度词、标点符号,如"患有、的、伴有"等.然后,将剩下的词替换成其对应的标准名称,由此分别得到有效元素集合 $setD$ 和 $setX$.对于 D_i,X_2 的有效元素集合 $setD$ 和 $setX$ 中的元素,本文迭代地用其上位词替换下位疾病成分以检测上位关系,直到出现以下这 2 种情况:若 $setX$ 中包含 $setD$,则 D_i 是 X_2 的上位词,并返回替换次数;否则,继续进行上位词替换,直到没有上位词可替换为止.最后,设定 X_3 为满足上位词条件且替换次数最少的 D_j.算法伪代码如算法 1 所示.

算法 1. 基于疾病构成成分的规则算法.

输入:ICD10 中标准疾病术语 X_1、临床诊断疾病数据中的常用语 X_2、疾病构成成分词典中的同义关系集合 R、停用词集合 $S=\{S_1, S_2, \cdots, S_n\}$、疾病构成成分词典中的上位关系 $HypernymMap$;

输出:疾病对$\langle X_1, X_2\rangle$的关系.

① 对X_1,X_2 根据双向最大匹配算法进行分词,分别得到 $X_1=\{X_{11}, X_{12}, \cdots, X_{1m}\}$,$X_2=\{X_{21}, X_{22}, \cdots, X_{2n}\}$的组成部分;
② for $X_{2i}\in X_2$ do
③ if $X_{2i}=S_i$ then
④ 将 X_{2i}移出 X_2;
⑤ else if X_{2i} in R then
⑥ 将 X_{2i}用在 R 中的标准同义词替换;
⑦ end if
⑧ end for
⑨ 对 X_1 同样进行步骤②~⑧操作;
⑩ 分别得到 X_2 的有效成分集 $setX$ 和 X_1 的有效成分集 $setD$;
⑪ if $setX-setD=\varnothing$ then
⑫ return 同义关系;
⑬ else if $setD\in setX$ then
⑭ return 上下位关系;
⑮ else while X_{2i} in $setX$ has hypernym in $HypernymMap$ do
⑯ 用 X_{2i}对应的 hypernym 替换 X_{2i};
⑰ if $setX-setD=\varnothing$ then
⑱ return 同义关系;
⑲ break;
⑳ else if $setD\in setX$ then
㉑ return 上下位关系;
㉒ break;
㉓ end if
㉔ end while
㉕ return 无关系;
㉖ end if

2.3.2 基于数据增强的 BERT 上下位关系识别算法

判别疾病医疗实体语义关系的问题,可看作一个分类任务,即 ICD10 中标准疾病术语 X_1 是否是常用语 X_2 的上位词.模型架构如图 2 所示:

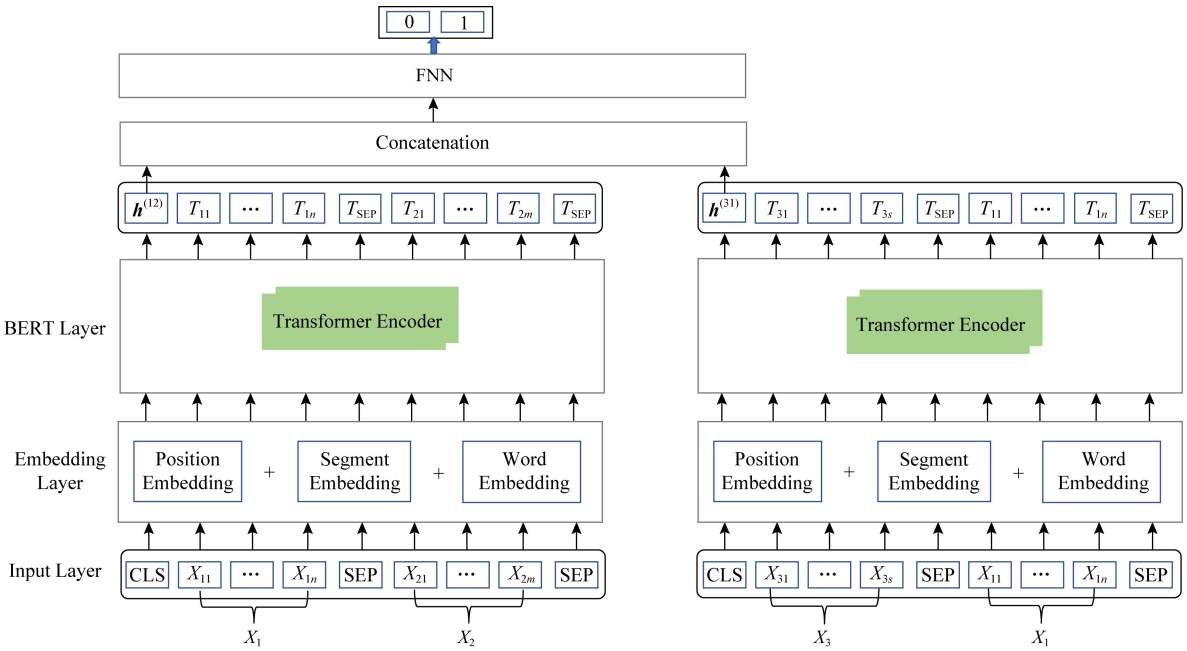

Fig. 2 Model of hypernym recognition algorithms of BERT based on data enhancement

图 2 基于数据增强的 BERT 上下位关系识别算法模型

本文利用预训练语言模型 BERT 对疾病对〈X_1,X_2〉和参考对〈X_3,X_1〉分别进行编码.以疾病对〈X_1,X_2〉为例,利用[SEP]标记标识 2 个疾病词的分割信息,在输入序列的开始位置加一个特殊的标记[CLS],拼接成"[CLS]X_1[SEP]X_2[SEP]"的形式作为输入.模型首先计算输入嵌入,输入嵌入包括词嵌入、句子嵌入和位置嵌入三者的总和.再将输入嵌入送入双向 Transformer 模型中,此时输出的[CLS]蕴含了 2 个疾病词是否相关的信息,我们使用标记[CLS]的最终输出 $\boldsymbol{h}^{(12)}$ 作为分类任务中的相关性表示向量.同理,将参考对〈X_3,X_1〉送入 BERT 后得到 $\boldsymbol{h}^{(31)}$.最后,将[$\boldsymbol{h}^{(12)}$,$\boldsymbol{h}^{(31)}$]拼接后送入前馈神经网络中,得到输出结果 0/1(0 表示两者无关系,1 表示 X_1 是 X_2 的上下位关系).

2.3.3 术语图谱关系识别算法对比实验

本文验证了构建融合常用语的疾病术语图谱中所用算法的有效性,我们利用区域医疗平台中的疾病数据作为实验数据集.特别地,该数据集中疾病名称之间的同义关系较少,因此直接利用基于疾病构成成分的规则算法判断同义关系,故本文仅针对上下位关系进行对比实验.

本文选取了 4 种关系识别算法,与本文所使用的方法进行对比:

1) 字符串相似度算法(string similarity algorithm).首先,求解出每一个疾病对中 ICD10 中标准疾病术语 X_1 和常用语 X_2 的 Levenshtein 距离 $distance(X_1,X_2)$,Levenshtein 距离指 2 个字符串之间由一个转成另一个所需的最少编辑操作次数.若 $distance(X_1,X_2)$ 结果超过阈值,则认为〈X_1,X_2〉具有上下位关系;否则,无关系.本文设置的阈值为 0.8.

2) 动态距离损失模型(dynamic distance loss model).文献[11]将每个常用语 X_2 训练出一个下位词向量 $\boldsymbol{O}_{X_2}$ 和一个上位词向量 $\boldsymbol{E}_{X_2}$.每当 X_2 作为下位词出现的时候,使用 $\boldsymbol{O}_{X_2}$;作为上位词候选出现的时候,使用 $\boldsymbol{E}_{X_2}$.然后,利用监督语料训练 SVM 模型,并利用训练好的模型判断输入疾病对〈X_1,X_2〉是否是上位词对.

3) 基于疾病构成成分的规则算法(rule algorithm based on disease components).文献[15]将疾病对〈X_1,X_2〉先根据词典进行分词,将分词后的元素进行去停用词、标准化操作,若 X_1 的元素包含在 X_2 的元素中,那么 X_1 是 X_2 的上位词,否则,迭代地用 X_2 元素的上位词替换 X_2 的该元素.

4) BERT.文献[16]将疾病对〈X_1,X_2〉以"[CLS] X_1[SEP] X_2[SEP]"的形式输入到预训练模型 BERT 中,在 BERT 模型后接一个前馈神经网络进行二分类.

对于关系识别结果,本文的评价指标采用最常用的 *Precision*,*Recall*,*F1_score* 作为评测标准,评估结果计算公式为:

$$Precision=\frac{\text{Number of right relationships}}{\text{Total number of relationships}}\times 100\%,$$

$$Recall=\frac{\text{Number of right relationships}}{\text{Total number of relationships in standard results}}\times 100\%,$$

$$F1\_score=\frac{2\times Precision\times Recall}{Precision+Recall}\times 100\%.$$

表 2 展示了 5 种对比算法的 *Precision*,*Recall*,*F*1_*score*.与现有算法相比,本文使用的算法获得最好的 *F*1_*score* 值,其 *Precision*,*Recall*,*F*1_*score* 分别为 97.18%,93.94%,95.53%.对于基于规则的关系识别方法,它的 *Precision* 达到了 100%,*Recall* 却很低,这是因为该算法受限于词典的规模而覆盖不全,但其预测的结果置信度很高.这也是本文融合该算法提供辅助信息的原因.此外,我们发现,本文算法的 *F*1_*score* 值比单独使用 BERT 高了 0.92%,证明了基于数据增强的 BERT 上下位关系识别算法的有效性.

Table 2 Comparative Experimental Results

表 2 对比实验结果 %

Algorithm	*Precision*	*Recall*	*F*1_*score*
String Similarity Algorithm	**100**	7.27	13.56
Dynamic Distance Loss Model	74.26	90.91	81.74
Rule Algorithm Based on Disease Components	**100**	22.12	36.23
BERT	96.24	93.03	94.61
Ours	97.18	**93.94**	**95.53**

Note: The best results are in bold.

2.4 添加 ICD11 层次结构信息

借助于 ICD10 与 ICD11 官网发布的映射表①,本文将融合了常用语的 ICD10 结构中的所有类目层疾病链接到 ICD11 的层次结构上,以添加 ICD11 层次结构信息,得到更加细粒度的疾病层级结构,方便医生查看和筛选疾病.添加 ICD11 层次结构信息的原因在于:

1) ICD10 的 3 位类目码的层次结构过于平面,没能体现出疾病间的层级结构.如图 3 所示,在 ICD10 中"糖尿病"与"内分泌疾病"处于同一层级上,而"糖尿病"应属于"内分泌疾病",即"糖尿病"应位于"内分泌疾病"的下层层级.

2) 疾病分类日益精细化,ICD11 调整了分类轴心、改变分类层次、增加或细化分类单元,对 ICD10 原有的分类结构和分类知识进行修订与完善.但鉴于医疗机构近 10 年来都采用的 ICD10 作为疾病编码标准,因此,需利用 ICD10 与常用语先做融合再添加 ICD11 的层次结构信息.

ICD10 标准的分类编码首先是类目,类目下分亚目,亚目下分细目,共 3 个层次.本文将 ICD10 类目层疾病映射到 ICD11 任意层疾病上,发现 ICD10 类目层能映射上 91.26%的 ICD11 中的疾病,于是将 ICD10 中的类目层的疾病(共 2 047 个)映射到 ICD11 各层节点的结果如表 3 所示.

表 3 中共映射了 2 521 条,而 ICD10 类目层疾病共 2047 条,多出 474 条的原因在于有 213 条数据不唯一映射.例如,ICD10 类目层的"其他细菌性肠道感染"(编码 A04)被进一步拆分成了 ICD11 中的

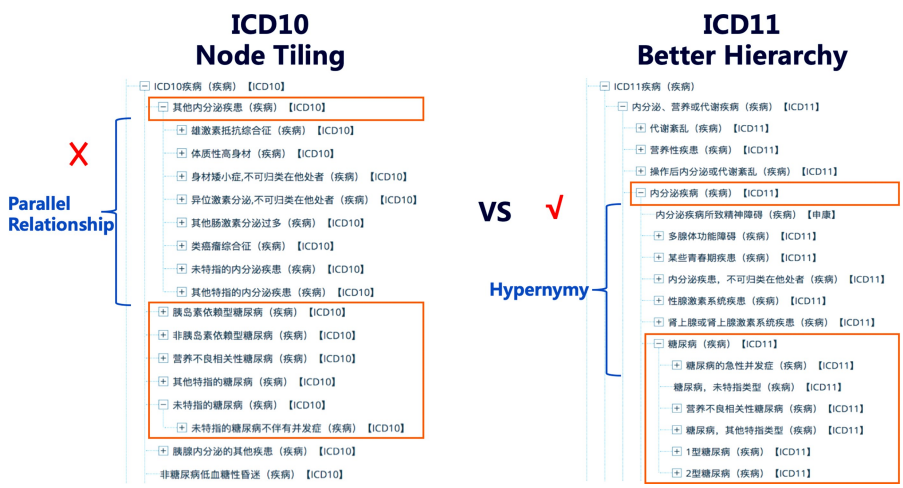

Fig. 3 Contrast schematic diagram of partial hierarchical structure between ICD10 and ICD11

图 3 ICD10 和 ICD11 部分层次结构对比示意图

① https://icd.who.int/browse11/l-m_st_infections/en

Table 3 Mapping ICD10 Category Layer to ICD11 Layers

表 3 ICD10 类目层映射到 ICD11 各层的映射

ICD11 Hierarchy	Mapping Number	Mapping Percentage/%
3rd	692	27.45
4th	1 161	46.05
5th	534	21.18
6th	120	4.76
7th	14	0.56

“其他弧菌的肠道感染”(编码 1A01)、“大肠埃希菌肠道感染”(编码 1A03)、“细菌性肠道感染、未特指的”(编码 1A02),从而导致映射不唯一.因此,需要把 ICD10 亚目层和细目层与 ICD11 的多重映射进一步对齐,而这需要专业医护人员介入,于是本文利用 2.5 节提到的基于疾病-科室关联图谱的任务分配方法对构建的融合常用语的大规模疾病术语图谱进行知识校验.

2.5 知识校验

即使经过数据增强,基于上文的上下位关系识别算法仍无法保证预测的上下位关系全部正确,可能出现 2 种错误情况:

1) 常用语与错误的 ICD10 名称具有关系.如“2 型糖尿病性神经病变”通过算法得出与“2 型糖尿病性神经炎”具有上下位关系,而正确的应该是“2 型糖尿病伴神经并发症”.

2) ICD10 名称不是常用语的直接上位词.本文将与常用语在层次结构上最相邻的上位词称为直接上位词.上下位关系具有传递性,即 X 是 Y 的直接上位词,Y 是 Z 的直接上位词,可得 X 是 Z 的上位词(非直接上位词).如“2 型糖尿病性大量白蛋白尿”通过算法与“2 型糖尿病”具有上下位关系,而“2 型糖尿病性肾病”才是“2 型糖尿病性大量白蛋白尿”的直接上位词.

上述情况的判断和纠正依赖于更深层的领域知识,为了确保疾病术语图谱的医学正确性,需要借助人工.

因此,本文设计了一种基于疾病-科室关联图谱的任务分配方法,方法流程如图 4 所示.首先获取 2.3.2 节算法预测出的所有疾病对〈ICD10 中的标准疾病术语,常用语〉集合生成待校验术语集,依据疾病对中标准疾病术语所对应的科室划分为多个基于科室的待校验术语子集.同一个待校验术语子集将被分配给同一科室的多名校对人员进行校验与修改,完成后交由机器自动进行一致性判断,校验结果

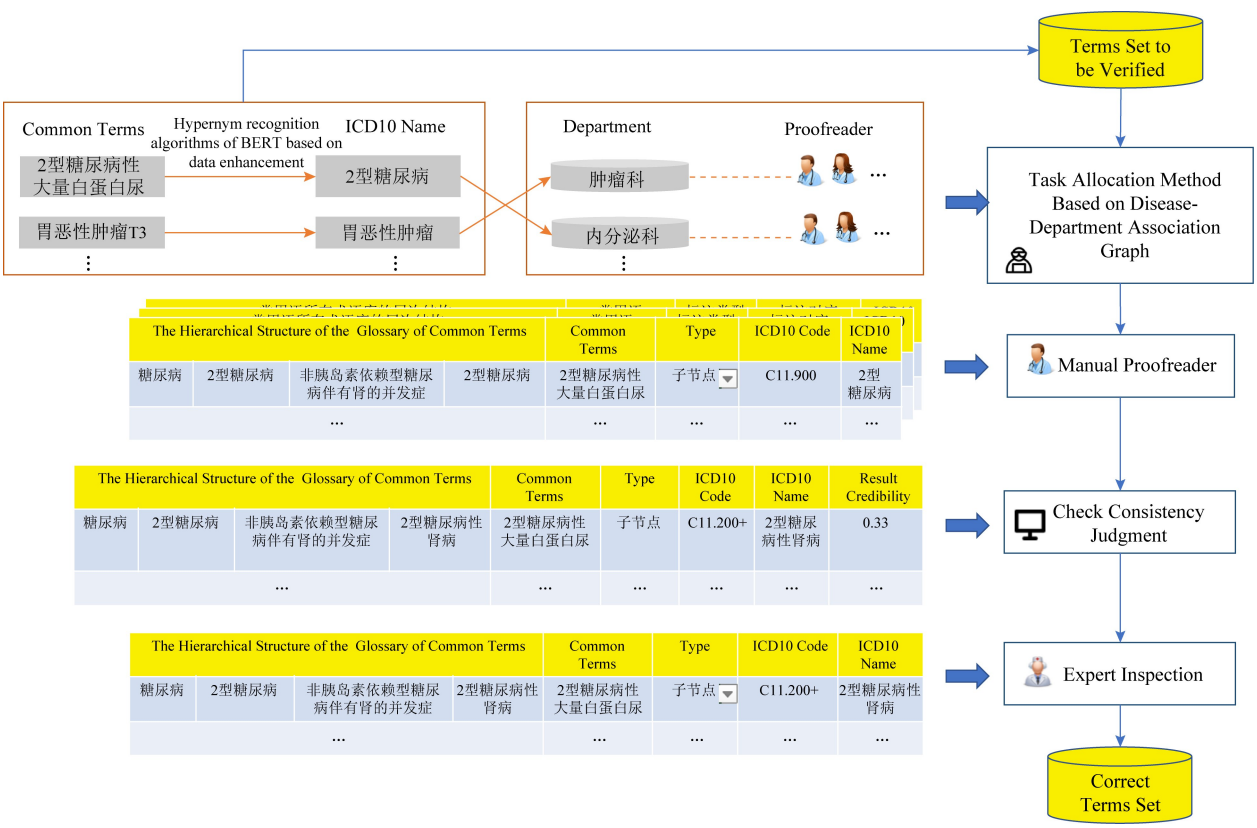

Fig. 4 Flow chart of task allocation method based on disease-department association graph

图 4 基于疾病-科室关联图谱的任务分配方法的流程图

可信度高于0.5的数据归为正确术语集,剩下的将由专家查验.

1) 基于疾病所在科室的任务分配.对于疾病对〈ICD10中的标准疾病术语,常用语〉,先根据所添加的ICD11的层次结构信息对疾病对中的标准疾病术语按照章节进行粗分类,再利用先前我们构造的疾病-科室知识图谱将每个章节下的标准疾病术语按科室进行细分类,最后归为同一科室的疾病对将填入同一个知识校验表格中,并扩展其标准疾病术语所在的层次结构.

2) 人工校对.同一任务将分配给 $n(n\geqslant 3)$ 名医护人员进行校验,目的是降低校验结果的随机性与偶然性.针对常用语链接上错误的ICD10名称的情况,医护人员修改图4知识校验表格中的[标注类型]或[标注对应ICD10编码];针对常用语链接上非直接上位词的情况,医护人员根据图4知识校验表格中[常用语所在术语库的层次结构]判断该常用语是否是直接上位词,并做相应修改.人工校对过程中对于某一条数据所有校对人员都没有修改的情况,则将该条数据直接加入正确术语集中.

3) 校对一致性判断.由于多人对同一条数据进行校对时,会出现多种修改情况.针对多人校对结果存在非一致性,需要对校对结果进行质量评估.

针对人工校对的结果,具体质量评估如下:将每条待校对数据视为一个校对任务 T_i,保证每个校验任务 T_i 都有 $n(n\geqslant 3)$ 名校对人员进行校验.每名校对人员在一次校验任务 D_i 中可能出现 m 种校对结果.因此,每个校对结果的可信度计算为

$$td_j=\frac{n_j}{n}(j=1,2,\cdots,m),$$

其中,n_j 表示选择第 j 种校对结果的人数.如果 $td_j>0.5$,则此次校对任务 T_i 的第 j 条校对任务结果正确,将其直接输出正确术语集;否则,将 T_i 交由医学专家进行查验.

4) 人机结合的方法节约了人力成本.首先,对于每一个常用语,算法都预测了在ICD10中的位置.虽然这个未必精确,但该所在的术语子树的分配一般来说是准确的.比如2.3.2节算法做错的"2型糖尿病性神经病变"与"2型糖尿病性大量白蛋白尿",都是术语"2型糖尿病"的子树.这样保证了数据搜索空间从ICD全集搜索缩减至子树搜索.而且糖尿病整体属于内分泌科,专科校对人员分配也是正确的,保障了人员可以对熟悉的疾病进行校验.

3 疾病术语编码评估

3.1 评估编码覆盖率

为了验证本文构建的疾病术语图谱能够有效覆盖更多临床诊断数据,我们从电子病历(electronic medical record, EMR)出院小结表中抽取出10 038条数据,作为第1组评估数据,并从随访数据中抽取出9 426条数据作为第2组评估数据,统计基于ICD10和本文构建的疾病术语图谱进行疾病编码能够映射成功的数量,映射结果如表4所示:

Table 4 Disease Names' Code Mapping

表4 疾病名称编码映射情况 %

Coding Mode	Mapping Rate of Data Coding	
	The 1st Group	The 2nd Group
Coding Based on ICD10	12.74	11.71
Coding Based on Our Method	97.19	77.87

从表4中可以看出使用了本文构建的疾病术语图谱相比于基于ICD10,编码覆盖率能平均提升75.31%,证明使用了疾病术语图谱能找到更多的疾病对应编码.但是本文构建的疾病术语图谱仍然没能找全所有疾病医学实体关系,其原因包含2个方面:1)由于真实数据出现2种疾病名称的情况.比如疾病名称"新生儿惊厥(癫痫)",其中"新生儿惊厥"在ICD10中对应编码是P90.x00,"癫痫"在ICD10中对应编码为G40.901,而"新生儿惊厥"和"癫痫"是2种疾病,疾病术语图谱难以根据算法分辨出其疾病编码.2)真实数据中出现并非疾病名称的数据,如"自体干细胞移植术后"、"后尿道瓣膜术后"等.对于第1种情况,可根据符号将包含2种编码的疾病名称设置不同的权重.对于第2种情况,出现非疾病名称的数据本不应该链接到疾病术语图谱上.

3.2 评估编码效率

为了验证本文构建的融合常用语的大规模疾病术语图谱在医疗领域中医生填写疾病编码时的优势,我们设置了人工编码和机器辅助编码2种评估方法,旨在对比构建的疾病术语图谱对医生编码疾病效率的影响.

对于人工编码,我们招募了5名熟悉ICD编码的医护人员,给定ICD10疾病标准分类编码,统计5名测试人员对随机采样的50条疾病名称找出匹配编码的完成时间.

对于机器辅助编码,我们首先利用本文构建的疾病术语图谱自动找出 50 条疾病名称对应的 ICD10 编码,以 2.5 节中图 4 所提及的知识校验表格的形式展示,然后 5 名校对人员对机器匹配结果进行校验.此时的完成时间定义为机器运行时间和校对人员校验所花费的时间之和.

实验结果如表 5 所示,使用了本文构建的疾病术语图谱辅助编码的完成速度是人工编码的 2.48 倍,表明了使用本文构建的疾病术语图谱自动进行疾病编码能够缩短医生的编码时间.实际情况下,医疗机构的医护人员对 ICD 编码体系并不是太熟悉,这也会影响编码效率,并且随着疾病数据量的增长,更能凸显将本文构建的疾病术语图谱应用于病案首页填写过程中的优势.

Table 5 The Completion Time Results of Manual Coding and Machine-Aided Coding

表 5 人工编码和机器辅助编码的完成时间结果

Proofreader	Manual Coding/s	Machine-Aided Coding/s
1	486.077	198.573
2	383.016	158.404
3	402.057	163.572
4	454.043	176.680
5	471.029	186.665
Average	439.244	176.779

3.3 评估编码正确率

利用区域平台电子健康记录(electronic health record, EHR)数据验证本文构建的疾病术语图谱的有效性,该数据包含上海 38 家三甲医院的挂号数据,含医生编码的数据占 536 456 条,对数据进行清洗后,随机抽取 2 个专病数据作为评估数据.本次评估目标在于统计医生手工编码和使用了本文构建的疾病术语图谱编码各自的正确率(accuracy),结果如表 6 所示.值得注意的是,评估数据的标准 ICD 编码以经过知识校验后得到的 ICD10 编码为准.

Table 6 Accuracy of Doctor's Manual Coding and Auto Coding by Disease Terminology Graph

表 6 医生手动编码和使用疾病术语图谱自动编码的正确率

Coding Mode	Accuracy/%
Doctor Manual Coding	19
Coding Based on Our Method	85

从表 6 结果可以看出,利用本文构建的疾病术语图谱编码的正确率远高于医生手动编码的正确率,提升了 66%.对医生手动编码正确率低的原因进行分析:1)医生对编码理解不统一,对于"2 型糖尿病性酮症"这一疾病名称,医生编码包括了 E11.103, E11.100, FFF 这 3 种编码,使用疾病术语图谱编码为 E11.100,经过校对人员校验,其与"2 型糖尿病性酮症酸中毒"为同义关系,编码应为 E11.100.2)部分医生填写疾病编码不规范.如常用语"胃恶性肿瘤Ⅳ期"应该链接在 ICD10 中"胃恶性肿瘤"(编码为 C16.900),而医生编码为 C16.再比如常用语"Ⅱ型糖尿病",其对应于 ICD10 中的"2 型糖尿病"(编码为 E11.900),而医生编码写作 E11.90000S.

4 结论与未来工作

本文通过基于疾病构成成分的规则算法和基于数据增强的 BERT 上下位关系识别算法辨别出常用语与 ICD10 中标准疾病术语的疾病医学实体关系,实现常用语与 ICD10 编码的映射,并且添加了 ICD11 的层级结构,方便医生查看疾病对应 ICD10 编码.利用本文构建的疾病术语图谱进行疾病编码在编码覆盖率、准确率以及编码效率 3 方面均有良好的表现.在未来,能够在各医疗结构中应用该疾病术语图谱,保障疾病编码的覆盖率、效率和准确率,推进医疗信息规范化进程.

参 考 文 献

[1] Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology [J]. Nucleic Acids Research, 2004, 32(Suppl1): D267-D270

[2] Fellbaum C. WordNet [M] //Theory and Applications of Ontology: Computer Applications. Berlin: Springer, 2010: 231-243

[3] Lenat D B. CYC: A large-scale investment in knowledge infrastructure [J]. Communications of the ACM, 1995, 38 (11): 33-38

[4] Hearst M A. Automatic acquisition of hyponyms from large text corpora [C] //Proc of the 14th Conf on Computational linguistics. Stroudsburg, PA: ACL, 1992: 539-545

[5] Navigli R, Velardi P, Faralli S. A graph-based algorithm for inducing lexical taxonomies from scratch [C] //Proc of the 22nd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2011:1872-1877

[6] Snow R, Jurafsky D, Ng A Y. Learning syntactic patterns for automatic hypernym discovery [C/OL] //Advances in Neural Information Processing Systems. 2005[2019-03-23]. http://papers.nips.cc/paper/2659-learning-syntactic-patterns-for-automatic-hypernym-discovery.pdf

[7] Yang Hui, Callan J. A metric-based framework for automatic taxonomy induction [C] //Proc of the 47th Conf of Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2009: 271-279

[8] Fu Ruiji, Guo Jiang, Qin Bing, et al. Learning semantic hierarchies via word embeddings [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2014: 1199-1209

[9] Tan Liling, Gupta R, van Genabith J. Usaar-wlv: Hypernym generation with deep neural nets [C] //Proc of the 9th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2015: 932-937

[10] Anh T L, Tay Y, Hui S C, et al. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network [C] //Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 403-413

[11] Yu Zheng, Wang Haixun, Lin Xuemin, et al. Learning term embeddings for hypernymy identification [C] //Proc of the 24th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 1390-1397

[12] Toral A, Monachini M. Named entity WordNet [C/OL] //Proc of the 6th Int Conf on Language Resources and Evaluation. 2008[2019-03-31]. http://ww w.lrec-conf.org/proceedings/lrec2008/pdf/188_paper.pdf

[13] Fellbaum C, Hahn U, Smith B. Towards new information resources for public health—From WordNet to MedicalWordNet [J]. Journal of Biomedical Informatics, 2006, 39(3): 321-332

[14] Vedula N, Nicholson P K, Ajwani D, et al. Enriching taxonomies with functional domain knowledge [C] //Proc of the 41st Int ACM SIGIR Conf on Research & Development in Information Retrieval. New York: ACM, 2018: 745-754

[15] Wang Qi, Wang Ting, Xu Chengming. Using a knowledge graph for hypernymy detection between Chinese symptoms [C] //Proc of the 10th Int Conf on Advanced Computational Intelligence (ICACI). Piscataway, NJ: IEEE, 2018: 601-606

[16] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint, arXiv: 1810.04805, 2018

Zhang Chentong, born in 1995. MSc. Student member of CCF. Her main research interests include information extraction and knowledge graph.

Zhang Jiaying, born in 1996. MSc. Student member of CCF. Her main research interests include information extraction, natural language processing, and knowledge graph.

Zhang Zhixing, born in 1996. MSc. Student member of CCF. His main research interests include information extraction, natural language processing, and knowledge graph.

Ruan Tong, born in 1973. PhD. Professor. Member of CCF. Her main research interests include natural language processing, knowledge graph, and information extraction.

He Ping, born in 1975. PhD. Professor. Her main research interests include hospital digital construction.

Ge Xiaoling, born in 1971. Master. Associate professor of health management. Her main research interests include hospital information management and data analysis.