基于融合条目词嵌入和注意力机制的自动 ICD 编码

张虹科[1] 付振新[1] 任前平[2] 徐辉[2] 赵东岩[1] 严睿[1,†]

1. 北京大学计算机科学技术研究所，北京 100871；2. 生命奇点（北京）科技有限公司，北京 100080;
† 通信作者，E-mail: ruiyan@pku.edu.cn

摘要 人工对病案首页的主要诊断进行 ICD 编码是一项费时费力并且容易出错的工作，针对此现状，提出了一种基于融合条目词嵌入和注意力机制的深度学习模型，可以充分利用电子病案中的多种非结构化文本数据，对病案首页的主要诊断进行自动 ICD 编码。该模型首先对含有病案条目的文本进行融合条目的词嵌入，并通过关键词注意力，丰富词级别的类别表示，然后利用词语注意力，突出重点词语的作用，增强文本表示，最后通过全连接神经网络分类器进行分类，输出 ICD 编码。在中文电子病案数据集上进行的消融实验验证了融合条目词嵌入、关键词注意力和词语注意力的有效性，并与多个基准模型相比，所提模型在对 81 种疾病的分类中取得了最好的分类效果，说明了所提模型可以有效提高自动 ICD 编码的质量。

关键词 自动 ICD 编码；融合条目词嵌入；关键词注意力；词语注意力；病案首页；主要诊断

中图分类号

Automated ICD Coding Based on Word Embedding with Entry Embedding and Attention Mechanism

ZHANG Hongke[1]，FU Zhenxin[1]，REN Qianping[2]，XU Hui[2]，ZHAO Dongyan[1]，YAN Rui[1,†]

1. Institute of Computer Science and Technology of Peking University, Beijing 100871; 2. Gennlife (Beijing) Technology Ltd, Beijing 100080; †Corresponding author, E-mail: ruiyan@pku.edu.cn

Abstract Manually ICD coding for the main diagnosis of the medical record home page is a time-consuming and error-prone task. In response to this situation, the author proposes a neural model based on word embedding with entry embedding and attention mechanism, which can make full use of the unstructured text in the electronic medical record to achieve automated ICD coding for the main diagnosis of the medical record home page. This method first embeds the words which contain the medical record entries into word embeddings, and enriches word-level representation based on keyword attention. Then, using the word attention to highlight the role of key words and enhance the text representation. Finally, outputs ICD codes by a fully connected neural network classifier. Ablation study on a Chinese electronic medical record data set shows that word embedding with entry embedding, keyword attention and word attention are effective. And the proposed model gets the best results for 81 diseases classification compared with baselines, which shows that the proposed model can effectively improve the quality of automated ICD coding.

Key words automated ICD coding; word embedding with entry embedding; keyword attention; word attention; medical record home page; main diagnosis

国际疾病分类(International Classification of Disease，ICD)是世界卫生组织(World Health Organization，WHO)制定的国际统一的疾病分类方法[1]。它根据疾病的病因、病理、临床表现和解剖位置等特性，将疾病分门别类，使其成为一个有序组合，并用编码的方法来表示疾病。为了各国能够统一使用，避免语言障碍导致编码不一致，ICD 编码统一采用英文字母加数字组合的形式，如 C22.901(肝恶性肿瘤)。目前广泛使用的是 ICD-10，全称为“疾病和有关健康问题国际统计分类”[2]。

国家重点研发计划(2017YFC0804001)，国家自然科学基金（61672058，61876196）

病案是医院对患者病情记录的详细文件，记录了患者在诊疗过程中的病情以及最终的诊疗结果[3]。病案首页是整个病案信息最核心、最集中的部分，记录了患者的基本信息、疾病诊断、住院天数、医疗费用等信息，是病案信息的综合反映。病案首页中的疾病诊断分为主要诊断和其他诊断，主要诊断即患者本次就医的目的，首要治疗的疾病；其他诊断指患者本次也进行治疗的疾病但不是最严重的，或者以往就有的疾病。对病案首页的主要诊断进行 ICD 编码是病案信息管理的核心技术，也是医疗付款中的疾病诊断相关分组(Diagnosis Related Groups，DRGs)[4]的基础和依据。很多医疗机构因为 ICD 编码的质量不如人意，造成了大量医疗付费的损失，同时也对患者的后续治疗产生了一定的影响。

对病案首页的主要诊断进行 ICD 编码并非易事。首先，ICD 编码规则相对复杂，而且种类繁多，有些编码表示的疾病非常相似，不容易区别。第二，住院患者往往存在有多种其他诊断，这些其他诊断会干扰主要诊断的判断，进而影响主要诊断的 ICD 编码。第三，医生在书写病案时因为个人习惯或者为了提升书写效率，往往会使用缩写或同义词，导致临床诊断不够清晰明确，给 ICD 编码人员带来了很大的困扰。因此，人工对病案首页的主要诊断进行 ICD 编码是一项费时费力并且容易出错的工作，如果能够对电子病案实现自动 ICD 编码，不仅可以辅助病案编码员编码，提高编码效率和质量，还可以对已编码的病案进行自动核查和纠错，对提高病案管理效率、降低病案管理成本和减少医疗付费损失有着十分重要的意义。

基于上述研究背景，本文提出了一种基于融合条目词嵌入和注意力机制的深度学习模型，可以充分利用电子病案中的多种非结构化文本数据，对病案首页的主要诊断实现高质量的自动 ICD 编码。与之前的研究相比，本研究的主要贡献有以下 3 点：

(1) 之前的研究大部分集中在单一文本建模，没有充分利用电子病案中的多种文本数据，丢掉了部分有用信息。针对这个问题，我们选取了多种文本进行建模，通过增加文本内容，丰富和强化文本的类别特征，可以更加全面地捕捉重要信息，提高分类的效果。

(2) 在词嵌入层，我们引入了病案条目信息，通过融合条目的词嵌入方法，丰富词语的分布式表示，可以体现同一词语在不同文本中的语义差异。

(3) 在词语表示层，我们引入了类别关键词信息，通过关键词注意力，形成基于关键词注意力分布的类别表示，进一步丰富词语的词级别表示。

1 相关工作

对电子病案进行自动 ICD 编码一直是研究的热点[5]。在早期，Larkey 和 Croft [6]通过组合 3 种分类器：K-近邻(K-nearest-neighbor)，关联反馈(relevance feedback)和贝叶斯独立分类器(Bayesian independence classifier)，对住院患者的出院记录实现了自动 ICD-9 编码。Franz 等[7]通过比较三种面向诊断短语的自动 ICD 编码方法，发现基于三元语言模型(Trigram Language Model)的检索方法比基于医疗词典的检索方法更有利于自动 ICD 编码。Kavuluru 等[8]通过对电子病历进行特征提取和特征选择，结合排序算法，实现了多标签的自动 ICD-9 编码。Koopma 等[9]基于规则并结合支持向量机(Support Vector Machines, SVM)提出了一个文本分类模型，可对癌症病人的死亡证明进行自动 ICD 编码。还有一些学者针对射线检查报告探索了多标签的自动 ICD-9-CM 编码[10]。早期的研究大多数都是基于规则或者特征工程的方法来对文本建模，没有考虑词语的上下文关系，忽略了重要词语与句子的贡献，无法很好地对内容丰富的电子病案进行表示。

近年来，随着深度学习的兴起与发展，一些学者将深度学习的技术应用到了自动 ICD 编码中[11]。Scheurwegs 等[12]基于分布式语义表示的思想，对未标注的语料提出了一种非监督的医疗实体抽取方法，可用于辅助自动 ICD 编码。Duarte 等[13]利用门控循环单元(Gated Recurrent Unit, GRU)和注意力机制(attention mechanism)，对癌症病人的死亡证明实现了自动 ICD-10 编码。Mullenbach 等[14]则基于卷积神经网络(Convolutional Neural Networks, CNN)和标签注意力(label attention)，对患者的出院记录实现了自动 ICD 编码。Shi 等[15]利用长短期记忆网络(Long Short-Term Memory, LSTM)和注意力机制对诊断描述建模，实现了多标签的自动 ICD 编码；在此基础上，Xie 等[16]则更进一步利用了一个序列树长短期记忆网络(tree-of-sequence LSTM)来表示 ICD 编码的层级结构，并提出了一个对抗学习算法来缓解不同医生书写病历的风格差异，并结合排序算法对多个标签进行排序，显著提高了多标签自动 ICD 编码的质量。Baumel 等[17]基于

两层的双向门控循环单元(Bidirectional Gated Recurrent Unit, BiGRU)分别在句子和文档两个层面进行表示，并加入了句子注意力(sentence attention)和标签注意力(label attention)来增强句子和文本表示，对于长文本的自动 ICD 编码有着不错的效果。Xu 等[18]充分利用了电子病历中的各种数据资源，结合机器学习和深度学习的优势，对不同类型的数据(非结构化、半结构化和结构化的数据)学习不同的分类器，然后集成起来，实现了基于多模态的自动 ICD 编码系统。由于深度学习在文本建模上具有的强大表征能力，不仅可以更好的表示词语与文本，还可以学习到词语的上下文关系和重要词语的信息，在文本分类领域展现出了强大的优势，因此，深度学习成为了目前研究自动 ICD 编码的主流方法。

2 中文电子病案数据集

本次研究所用的中文电子病案数据集来源于山东省某医院近 5 年(2012 年至 2017 年)住院患者的真实电子病案记录。数据集总共包含了 82375 例住院患者的病案记录，这些记录包括结构化的数据和非结构化的数据，每类数据都有对应的病案条目，比如年龄、性别、主诉、现病史、出院记录等。由于数据内容丰富庞杂，为了方便研究，经过咨询病案编码员与临床医生，我们选取了 6 种对编码比较重要的非结构化文本数据进行建模，这 6 种文本对应的病案条目分别是主诉、现病史、首次病程记录、检查报告、查房记录和出院记录，表 1 展示了一例病案记录中这 6 种文本的简略内容以及对应的主要诊断和 ICD 编码。

表 1　一例病案记录的简略内容以及对应的主诊断和 ICD 编码

Table 1 A brief description of a medical record and corresponding main diagnosis with ICD code

病案条目	文本内容	主要诊断	ICD 编码
主诉	右上腹痛 1 周。	肝恶性肿瘤	C22.901
现病史	患者于 1 周前无明显诱因出现腹痛，…… 尿色深黄，呈浓茶色，体重无明显变化。		
首次病程记录	中年男性患者，右上腹痛 1 周。既往“糖尿病”病史 8 年，…… 谷丙转氨酶 44IU/L。		
检查报告	肝脏形态欠规整，左叶增大……脾脏形态明显增大，腹膜后见广泛迂曲血管影。		
查房记录	患者腹痛有所缓解，无恶心、呕吐。……保肝退黄、利胆治疗，继续观察病情变化。		
出院记录	患者入院后完善相关辅助检查，……建议院外继续保肝治疗，肿瘤内科、介入科随诊。		

3 模型方法

为了方便研究，本文的自动 ICD 编码问题可描述为，给定含有病案条目的文本数据X，输出对应的 ICD 编码Y，目标是找到一个最优映射F，使得$F: X \to Y$。对任意的 $x_i \in X$，可用式(1)表示：

$$x_i = \{k_1: [w_{11}^i, w_{12}^i, \dots, w_{1L_1}^i], k_2: [w_{21}^i, w_{22}^i, \dots, w_{2L_2}^i], \dots, k_m: [w_{m1}^i, w_{m2}^i, \dots, w_{mL_m}^i]\} \tag{1}$$

其中，k_j为第j个病案条目，$j \in \{1,2,\dots,m\}$，$[w_{j1}^i, w_{j2}^i, \dots, w_{jL_j}^i]$为第$i$个样本第$j$个病案条目的词语序列，$L_j$为序列长度。对应的 ICD 编码集为$Y = \{c_i\}$，$i = 1,2,\dots,n_c$，$n_c$为 ICD 编码类别数。

本文所提模型的基础框架为两层双向长短期记忆网络[19] (Bidirectional Long Short-Term Memory, BiLSTM)的文本编码器和一层全连接神经网络分类器，并在文本编码器中引入了三个模块：融合条目的词嵌入模块、关键词注意力模块和词语注意力模块。模型的整体框架如图 1 所示。

3.1 融合条目的词嵌入模块

电子病案中的病案条目说明了文本描述的类型和主题，反映了患者在某方面的情况。同一个词语，在不同条目的文本中，其意义和重要性可能有所不同，比如，在现病史中，对很多患者都会有这样的描述：“无咳嗽、咳痰、胸闷、胸痛”，表明该患者没有明显的心肺疾病，但如果在主诉中出现了“咳嗽、咳痰、胸闷、胸痛”这些词语，说明患者是有明显的心肺疾病的。因此，病案条目不仅可以给词语提供主题信息，还可以强化上下文的语境，反映了同一词语在不同文本中的差异。为了体现词语的这种差异，我们设计了融合条目的词嵌入模块，在对文本的词语进行词嵌入的同时也对相应的病案条目进行词嵌入，并将条目向量融入到词向量中，丰富词语的分布式表示。

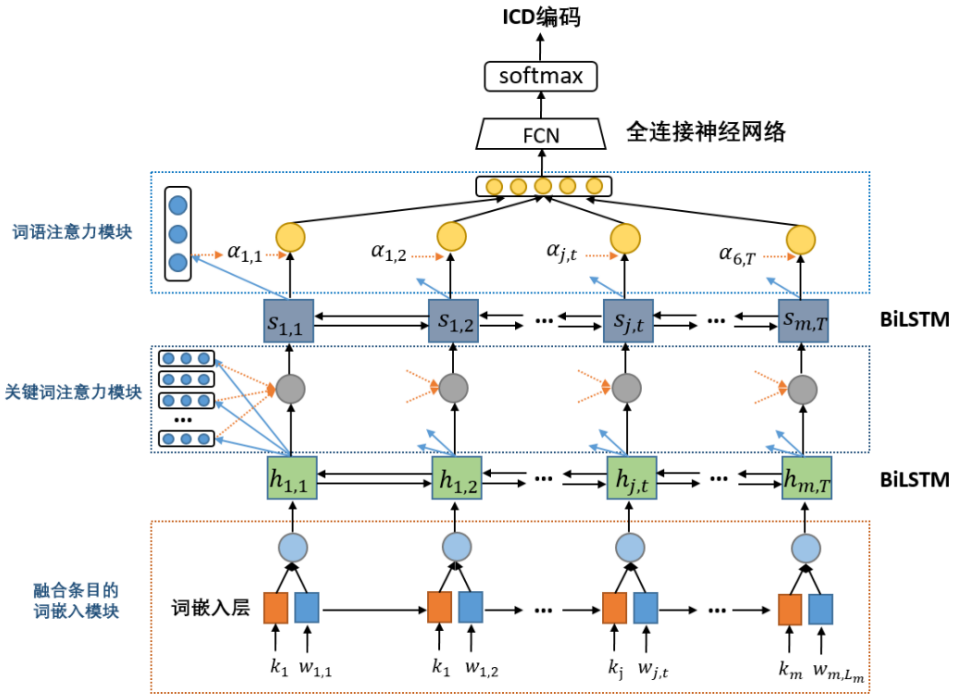

图 1　模型整体框架

Fig.1 Model framework

该模块包括两个独立的词嵌入层，分别对词语和对应的病案条目进行词嵌入，并将获得的两个词向量拼接起来作为融合条目的词向量。融合条目的词向量$\boldsymbol{v}_{j,t}$可用下列式子表示：

$$\boldsymbol{k}_j = f_{lookup}(k_j, \boldsymbol{W}_{emb}^{k}) \tag{2}$$

$$\boldsymbol{w}_{j,t} = f_{lookup}(w_{j,t}, \boldsymbol{W}_{emb}^{w}) \tag{3}$$

$$\boldsymbol{v}_{j,t} = [\boldsymbol{k}_j; \boldsymbol{w}_{j,t}] \tag{4}$$

其中，$\boldsymbol{W}_{emb}^{k}$和$\boldsymbol{W}_{emb}^{w}$分别是病案条目和普通词语对应的词嵌入矩阵，是模型需要学习的参数；$f_{lookup}()$是查表函数。融合条目的词向量$\boldsymbol{v}_{j,t}$将会输入第一层 BiLSTM 中进行编码，获得词语的初步表示。

3.2 关键词注意力模块

不同的 ICD 编码有不同的关键词，这些关键词对于分类模型来说是十分重要的信息，因此，我们在词语表示层引入了关键词信息，通过关键词注意力形成词级别的类别表示，进一步丰富词语的表示，并将该表示与分类模型结合起来，提升分类性能。该模块包括关键词提取和关键词注意力两个过程，如图 2 所示。

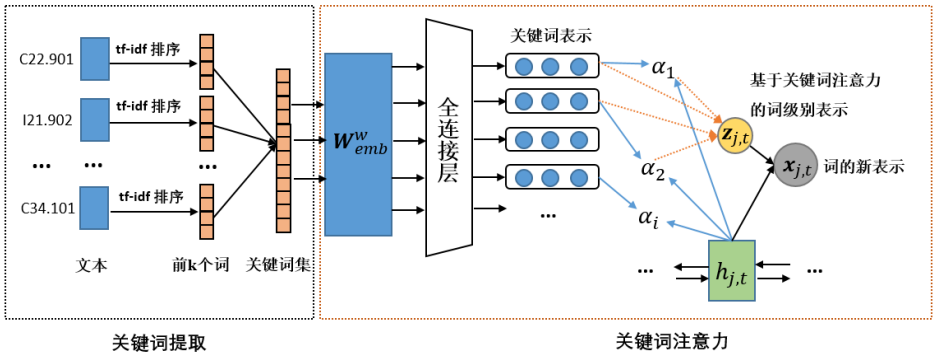

图 2　关键词注意力模块

Fig.2　Keyword attention module

3.2.1 基于 TF-IDF 算法的关键词提取

在进行关键词注意力之前，先基于 TF-IDF(Term Frequency-Inverse Document Frequency)[20]算法提取出各类 ICD 编码的关键词。首先，将所有文本合并成一类文本，然后计算每个词语在不同 ICD 编码中的 TF-

IDF 值并进行排序，最后针对每个类别，选取 TF-IDF 值最大的前 k 个的词语作为类别关键词。对于在某一类文档d_j里的词语t_i来说，它的 TF-IDF 计算公式为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{5}$$

$$idf_i = \log\left(\frac{|D|}{1+|\{j:t_i \in d_j\}|}\right) \tag{6}$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{7}$$

其中，$n_{i,j}$是该词在文档d_j中出现的次数，$\sum_k n_{k,j}$是文档d_j中所有词语出现的次数之和。$|D|$是语料库中的文档总数，$|\{j: t_i \in d_j\}|$为包含词语t_i的文档数目。表 2 列出了部分 ICD 编码的前 10 个关键词。

表 2　部分 ICD 编码的前 10 个关键词
Table 2　Top 10 keywordss for some ICD codes

ICD 编码	关键词
I25.101	胸闷 冠心病 双肺 病情 心脏 胸痛 发作 血压 未闻 呼吸
I61.005	基底节 脑出血 肢体 ct 颅脑 病情 肌力 破入 瞳孔 血压
C22.901	肝癌 腹部 肝脏 ct 占位 tce 增强 iu 保肝 肿瘤
C15.901	食管癌 食管 化疗 淋巴结 双肺 放疗 胃镜 总数 呼吸 辅助
C16.902	腹部 胃癌 淋巴结 化疗 切缘 不适 胃镜 胃窦 转移 病理

3.2.2 关键词注意力

将关键词提取出来之后，组成关键词集，然后经过普通的词嵌入矩阵$\boldsymbol{W}_{emb}^{w}$和一个全连接神经网络得到关键词的表示，再与第一层 BiLSTM 的输出(词语的初步表示)进行关键词注意力操作，获得每个词语基于关键词注意力分布的词级别表示。最后该表示与词语的初步表示拼接起来组成词语的新表示。词语的新表示$\boldsymbol{x}_{j,t}$由下列公式计算：

$$\boldsymbol{h}_{j,t} = BiLSTM_1(\boldsymbol{v}_{j,t}; \boldsymbol{W}_1) \tag{8}$$

$$\boldsymbol{e}_i = f_{lookup}(e_i, \boldsymbol{W}_{emb}^{w}) \tag{9}$$

$$\boldsymbol{u}_{\boldsymbol{e}_i} = tanh(\boldsymbol{W}_k \boldsymbol{e}_i + \boldsymbol{b}_k) \tag{10}$$

$$\alpha_{u_i} = \frac{exp(\boldsymbol{h}_{j,t}^{T} \boldsymbol{u}_{\boldsymbol{e}_i})}{\sum_l exp(\boldsymbol{h}_{j,t}^{T} \boldsymbol{u}_{\boldsymbol{e}_l})} \tag{11}$$

$$\boldsymbol{z}_{j,t} = \sum_i \alpha_{u_i} \boldsymbol{u}_{\boldsymbol{e}_i} \tag{12}$$

$$\boldsymbol{x}_{j,t} = [\boldsymbol{h}_{j,t}; \boldsymbol{z}_{j,t}] \tag{13}$$

其中，$\boldsymbol{W}_1$是第一层 BiLSTM 的参数，$\boldsymbol{h}_{j,t}$是词语的初步表示；$\boldsymbol{W}_k$和$\boldsymbol{b}_k$是全连接层的参数，$\boldsymbol{u}_{\boldsymbol{e}_i}$是关键词$e_i$的表示，$\boldsymbol{z}_{j,t}$是基于关键词注意力分布的词级别表示。

3.3 词语注意力模块

因为并不是所有的词语都对文本语义起同等大小的贡献，因此，基于 Yang 等[21]的研究，我们在第二层的 BiLSTM 中引入了词语注意力模块，获取每个词语对文本表示的重要程度，以突出重点词语的贡献，弱化无关词语的影响，可以更好地表示文本。具体地，先通过一个全连接层得到隐状态$\boldsymbol{s}_{j,t}$的表示$\boldsymbol{v}_{s_{j,t}}$

$$\boldsymbol{s}_{j,t} = BiLSTM_2(\boldsymbol{x}_{j,t}; \boldsymbol{W}_2) \tag{14}$$

$$\boldsymbol{v}_{s_{j,t}} = tanh(\boldsymbol{W}_w \boldsymbol{s}_{j,t} + \boldsymbol{b}_w) \tag{15}$$

其中，$\boldsymbol{W}_2$是第二层 BiLSTM 的参数，$\boldsymbol{W}_w$和$\boldsymbol{b}_w$是全连接层的参数。然后，计算单词的注意力权重$\alpha_{j,t}$

$$\alpha_{j,t} = \frac{exp(\boldsymbol{v}_{s_{j,t}} \boldsymbol{v}_w)}{\sum_{m,n} exp(\boldsymbol{v}_{s_{m,n}} \boldsymbol{v}_w)} \tag{16}$$

词语上下文向量$\boldsymbol{v}_w$经过随机初始化后，在训练期间与模型参数共同学习更新。最后，整合词语的注意力权重$\alpha_{j,t}$，得到文本表示$\boldsymbol{d}$

$$\boldsymbol{d} = \sum_{j,t} \alpha_{j,t}\, \boldsymbol{s}_{j,t} \tag{17}$$

3.4 全连接神经网络分类器

获得文本表示$\boldsymbol{d}$之后，经过全连接神经网络和 softmax 函数，得到 ICD 编码的类别预测值$\hat{y}$，并通过最小化交叉熵损失函数，逐层更新网络参数，可获得一个文本分类器，实现自动 ICD 编码。该过程可由以下式子表示：

$$\hat{y} = softmax(\boldsymbol{W}_c\boldsymbol{d} + \boldsymbol{b}_c) \tag{18}$$

$$min \;\; \mathcal{L} = -[ylog(\hat{y}) + (1-y)log(1-\hat{y})] \tag{19}$$

其中，$\boldsymbol{W}_c$和$\boldsymbol{b}_c$是全连接神经网络的参数，$\mathcal{L}$为交叉熵损失函数。

4 实验与结果

4.1 数据预处理

首先，从数据集中抽取出选定的 6 种文本数据和对应的 ICD 编码。由于数据集中包含的 ICD 编码较多，约有 640 种，其中有很多是不常见的疾病编码(比如 M34.902，硬皮病)，其对应的病例数很少，不方便展开研究。因此，我们把病例数小于 100 例的 ICD 编码过滤掉，最终保留了 81 种常见疾病的 ICD 编码，共 75431 例数据，其中训练集 60345 例(80%)，验证集 7543 例(10%)，测试集 7543 例(10%)。图 3 展示了训练集中病例数最多的前 25 种 ICD 编码的分布情况。

从图 3 可以看出，ICD 编码的分布不平衡，因此在训练的时候，针对病例数少于 500 例的 ICD 编码，我们进行了随机重采样，保证每类 ICD 编码的病例数在 500 例以上。对文本数据先用 jieba 中文分词工具进行分词操作，然后过滤掉数字、标点和低频词(词频低于 5 的词语)。

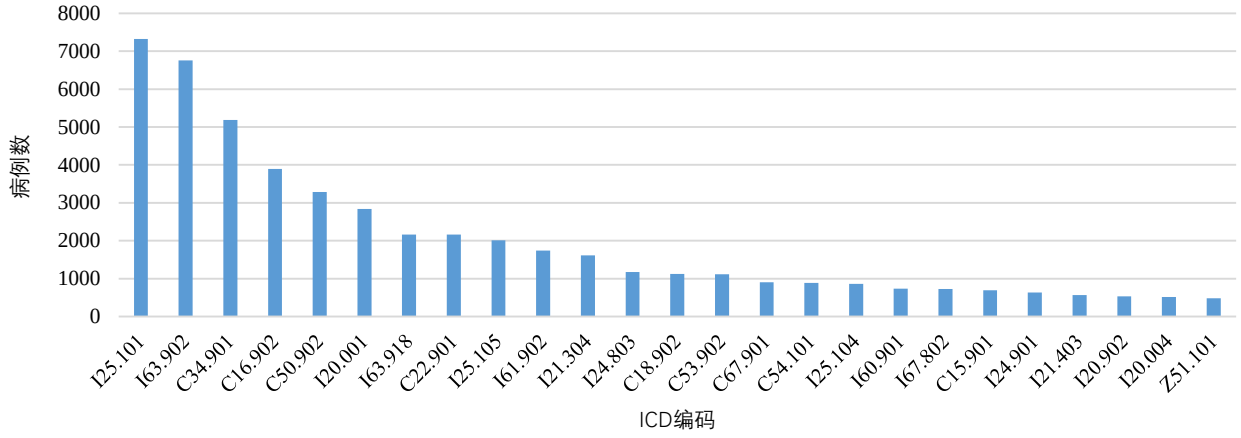

图 3　训练集中部分 ICD 编码的分布情况

Fig.3　Distribution of partial ICD codes in training set

4.2 实验设置

病案条目词向量的维度设为 50，普通词向量维度设为 128，两层 BiLSTM 的神经元数都设为 128，每类 ICD 编码的关键词数设为 10，损失函数采用 Adam 优化器优化，学习率设为 0.001。

为了体现各个模块的作用，我们进行了消融实验，通过去掉某个模块前后的效果比较，验证这个模块的作用。No-EE(No-Entry Embedding)表示去掉融合条目词嵌入模块之后的模型；No-KA(No-Keyword Attention)表示去掉关键词注意力模块之后的模型；No-WA(No-Word Attention)表示去掉词语注意力模块之后的模型；BiLSTM 表示去掉所有模块之后，仅包含两层 BiLSTM 和全连接层的模型。

除此之外，我们还设置了几个基准模型：基于 TF-IDF 特征表示的支持向量机[22] (SVM)模型和随机森林[23](Random Forest)模型、基于卷积神经网络的 Text-CNN[24]文本分类模型以及基于双层注意力网络的自动 ICD 编码模型 HA-GRU[17]。

针对多分类问题，我们采用了准确率(Accuracy)和经过宏平均(Macro-averages)的精确率(Precision)、召回率(Recall)、F1 分数(F1-score)作为模型的评价指标。

4.3 实验结果

实验结果见表 3。

从消融实验的结果可以看出，融合条目的词嵌入模块能够提升 1.27%的准确率和 3.5%的 F1 分数，对精确率和召回率也都有小幅度的提升，说明了通过融合条目的词嵌入方法来丰富词语的分布式表示，确实能够增强分类效果，提升自动 ICD 编码的质量。关键词注意力模块在准确率上提升了 2.06%，在 F1 分数上提升了 5.61%，说明了基于关键词注意力形成的词级别表示，能够利用关键词的类别信息，有助于提升文本分类的性能。词语注意力模块的提升效果最为明显，在准确率上提升了接近 10%，在精确率、召回率和 F1 分数上都有大幅度的提升(20%以上)，说明了在文本建模方面，特别是对长文本分类来说，词语注意力的重要性是毋庸置疑的。与不加三个模块的 BiLSTM 相比，加了两个模块和三个模块的模型在准确率、精确率、召回率和 F1 分数上都有不同幅度的提升，更进一步说明了所提模块的作用。

从与基准模型的结果比较可以看出，我们提出的模型明显优于传统的基于 TF-IDF 特征表示的支持向量机模型和随机森林模型，体现了深度学习的优势。与基于卷积神经网络的 text-CNN 文本分类模型相比，我们的模型在准确率上高出了 10.5%，在精确率、召回率和 F1 分数上的提升更为明显，体现了循环神经网络结合注意力机制在文本分类方面的优势。虽然我们的模型和 HA-GRU 模型都是基于循环神经网络和注意力机制，但我们在词向量中融入了病案条目信息，并在词语表示层引入了关键词注意力，使得词语的表示更为丰富。从实验结果来看，相比于 HA-GRU 模型，我们的模型在准确率上提升了 1.09%，在 F1 分数上提升了 12.61%，在精确率和召回率上也有明显的提升，说明了我们的模型更优。

表 3　实验结果

Table 3　Experimental result

实验	模型	Accuracy	Macro-averages		
			Precision	Recall	F1-Score
基准模型	SVM	0.6235	0.4399	0.4479	0.4385
	Random Forest	0.6939	0.4158	0.3384	0.3372
	Text-CNN	0.7152	0.5023	0.3339	0.3575
	HA-GRU	0.8095	0.5806	0.5247	0.5223
消融实验	BiLSTM	0.6910	0.5009	0.4156	0.4542
	No-EE	0.8077	0.6916	0.6002	0.6134
	No-KA	0.7998	0.6885	0.5665	0.5923
	No-WA	0.7386	0.5264	0.4017	0.4556
	Our Full Model	**0.8204**	**0.7243**	**0.6429**	**0.6484**

说明: 粗体数字表示最好的结果

5　结论

本文通过融合条目词嵌入和关键词注意力，丰富了词语的表示，并结合词语注意力，对中文电子病案中的主诉、现病史、首次病程记录、检查报告、查房记录和出院记录 6 种非结构化的文本进行建模，学习出一个基于深度学习的文本分类模型，可以对病案首页中的主要诊断进行自动 ICD 编码。通过消融实验和与基准模型的对比实验，验证了所提模型的有效性。

在接下来的工作中，我们将会考虑将更多的文本数据和结构化的数据加入到模型中，以进一步提升自动 ICD 编码的质量。由于本次研究的疾病种类还比较窄，涉及的 ICD 编码比较少，在今后的研究中，我们将会考虑增加对非常见病的建模，以扩大自动 ICD 编码的范围。

参考文献

[1] World Health Organization. International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index[J]. 1978.

[2] Uribe M O. International Classification of Diseases, World Health Organization, Tenth version ICD-10[J]. Salud Mental, 1996, 19: 11-18.

[3] 卫生部. 病历书写基本规范(试行)[J]. 中国卫生法制, 2002, 1(5):183-186.

[4] Horn S D, Bulkley G, Sharkey P D, et al. Interhospital differences in severity of illness: problems for prospective payment based on diagnosis-related groups (DRGs)[J]. New England Journal of Medicine, 1985, 313(1): 20-24.

[5] Stanfill M H, Williams M, Fenton S H, et al. A systematic literature review of automated clinical coding and classification systems[J]. Journal of the American Medical Informatics Association, 2010, 17(6): 646-651.

[6] Larkey L S, Croft W B. Combining classifiers in text categorization[C]//SIGIR. 1996, 96: 289-297.

[7] Franz P, Zaiss A, Schulz S, et al. Automated coding of diagnoses--three methods compared[C]//Proceedings of the AMIA Symposium. American Medical Informatics Association, 2000: 250.

[8] Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records[J]. Artificial intelligence in medicine, 2015, 65(2): 155-166.

[9] Koopman B, Zuccon G, Nguyen A, et al. Automatic ICD-10 classification of cancers from free-text death certificates[J]. International journal of medical informatics, 2015, 84(11): 956-965.

[10] Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems[C]//BMC bioinformatics. BioMed Central, 2008, 9(3): S10.

[11] Shickel B, Tighe P J, Bihorac A, et al. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis[J]. IEEE journal of biomedical and health informatics, 2018, 22(5): 1589-1604.

[12] Scheurwegs E, Luyckx K, Luyten L, et al. Assigning clinical codes with data-driven concept representation on Dutch clinical free text[J]. Journal of biomedical informatics, 2017, 69: 118-127.

[13] Duarte F, Martins B, Pinto C S, et al. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text[J]. Journal of biomedical informatics, 2018, 80: 64-77.

[14] Mullenbach J, Wiegreffe S, Duke J, et al. Explainable prediction of medical codes from clinical text[J]. arXiv preprint arXiv:1802.05695, 2018.

[15] Shi H, Xie P, Hu Z, et al. Towards automated icd coding using deep learning[J]. arXiv preprint arXiv:1711.04075, 2017.

[16] Xie P, Xing E. A neural architecture for automated icd coding[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1066-1076.

[17] Baumel T, Nassour-Kassis J, Cohen R, et al. Multi-label classification of patient notes: case study on ICD code assignment[C]//Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[18] Xu K, Lam M, Pang J, et al. Multimodal Machine Learning for Automated ICD Coding[J]. arXiv preprint arXiv:1810.13348, 2018.

[19] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[20] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

[21] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1480-1489.

[22] Joachims T. Text categorization with support vector machines: Learning with many relevant features[C]//European conference on machine learning. Springer, Berlin, Heidelberg, 1998: 137-142.

[23] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. Journal of chemical information and computer sciences, 2003, 43(6): 1947-1958.

[24] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.