文章编号：2095-2163(2019)01-0114-05 中图分类号：TP391.1 文献标志码：A

基于深度学习的煤矿领域实体关系抽取研究

杜 嘉，刘思含，李文浩，徐啸迪，刘旭红
（北京信息科技大学 计算机学院，北京 100101）

摘 要：关系抽取是构建知识图谱的一个重要过程。为了更好地构建煤矿领域知识图谱，本文对关系抽取的方法进行研究。传统关系抽取方法在训练前多需要人工选取特征、大量标注数据、且需要专业领域的专家辅助、费时费力、且成本较高。本文采用字向量和深度学习相结合的方法对实体间的关系进行抽取，降低数据标注的难度，提高训练效率。实验结果证明使用字向量与深度学习相结合的方法能够较有效地完成煤矿领域实体关系抽取的任务。

关键词：关系抽取；知识图谱；循环神经网络；字向量

Research on entity relationship extraction in coal mine based on deep learning

DU Jia, LIU Sihan, LI Wenhao, XU Xiaodi, LIU Xuhong
(Computer School, Beijing Informatin Science and Technology University, Beijing 100192, China)

【Abstract】 Relational extraction is an important process in building knowledge maps. In order to better construct the knowledge map of coal mine field, this paper studies the method of relationship extraction. In this paper, the combination of word vector and deep learning is used to extract the relationship between entities, which reduces the difficulty of data annotation and improves training efficiency. The experimental results prove that the combination of word vector and deep learning can effectively complete the task of extracting entity relations in coal mine field.

【Key words】 relational extraction; knowledge maps; RNN; character embedding

0 引 言

煤炭是中国的重要基础能源，在使用的能源中占有极大比例。近年来，煤矿领域的信息量与日俱增，关于煤矿事故的信息也逐渐积累。在大量的煤矿事故中，含有煤矿事故发生的原因、解决的方案、责任的落实等重要信息，这类信息可为预防煤矿事故再次发生提供参数、分析、教训的有效信息。但是由于煤矿领域信息多为无结构文本，难以直接进行有效利用，因此需要以一种更加有效的方式对煤矿领域，煤矿事故的信息进行描述，并进一步分析。

知识图谱采用可视化技术来展示信息的发展、结构关系，同时对信息之间联系进行描绘，极大方便了对信息的分析、利用。

信息抽取是构建知识图谱的重要步骤。目前信息抽取是自然语言处理领域的一个重要分支，其任务是从自然文本中抽取出结构化的信息。关系抽取是信息抽取研究中的重要内容，是构建知识图谱的重要步骤，实体关系抽取的准确率将极大影响所构建的知识图谱的质量。因此，研究关系抽取问题对构建知识图谱有着积极意义。

1 关系抽取研究现状

关系抽取作为信息抽取的重要节点，一直都是国内外研究的一个重要方向。在关系抽取领域，方法繁多。早期关系抽取仅从模式匹配、词典等方向去分析。之后不断发展，引入了机器学习方法、深度学习方法，是目前研究的重点。文献[1]提出标注传播算法，对关系进行聚类抽取，但噪声对最终的准确率有所影响。2013 年，邵堃等[7]采用模式匹配的方法抽取结构化信息，采用动态模式库以提高抽取的准确率，但分词的结构，专业词汇的存在都会影响到识别的效果。

机器学习方法分为有监督方法、半监督方法、无监督方法等。有监督的机器学习方法一般将关系抽取看为一个分类问题。也就是对不同的实体对，在不同语句中的关系分类。一般需要提前定义关系的类别。例如条件随机场算法[3]（Conditional random Field，CRF）、支持向量机方法[4]（Support Vector Machine，SVM）、kNN 算法[5]等都是有监督的方法，目前被广泛应用于关系抽取领域，并且取得了出众的效果。Kambhatla 提出最大熵模型[6]（Maximum

基金项目：北京信息科技大学 2018 年人才培养质量提供经费(5111823402)。

作者简介：杜 嘉(1997-)，男，本科生，主要研究方向：自然语言处理、深度学习。

收稿日期：2018-10-22

Entropy Model，MEM)，结合语义特征，句法分析抽取关系。陈宇利用 DBN[7](Deepth Belief Network)证明在中文关系抽取领域，使用字特征进行关系抽取比使用词特征进行关系抽取效率更高。基于特征向量的方法则将文本信息转化为数字信息，用启发式方法选取特征，准确率较高。但新的特征越来越难以寻找，之后的算法效率的提高将更加困难。为了克服基于特征向量方法的缺陷，核函数概念被引入到机器学习中[8-9]。核函数方法采用字符串或者句法分析树作为算法的输入信息，通过计算输入信息之间的相似度实现分类效果。核函数方法解决了部分基于特征向量的方法所遇到的问题，提高了关系抽取的准确率。但基于核函数的方法计算复杂度较高，并且容易引入噪声，不适合从大规模的语料中抽取关系。半监督方法如自举方法减少了训练过程中对标注语料的依赖，降低了人工标注的成本，但存在语义漂移问题。无监督方法则主要使用聚类算法，能够应用于大规模开放性信息领域中，但是难以对关系名称进行准确描述。

为了提高关系抽取的质量、效率，本文采用字向量的方式表示文本数据，结合深度学习方法，用分类的方式抽取实体关系。字向量可以更好地表现出文本数据中字与字之间的内在联系，使用深度学习的方法，通过训练过程学习文本数据间的联系，完成分类任务。

2　煤矿领域实体关系抽取方法

2.1　实体关系抽取框架

实体关系抽取建立在实体抽取的基础上。煤矿领域实体关系的抽取问题最终转化为对已知文本中的实体对的情况下，根据文本内容对文本中的实体对进行分类的问题。本文采用字向量方法对文本以及文本中的实体对进行描述，采用双向循环神经网络加注意力机制[10]对实体对进行分类，实现煤矿领域关系抽取的方法。方法框架如图 1 所示。

具体步骤如下：

从煤矿安全网、煤矿事故网、安全管理网上爬去煤矿事故案例报道和分析报告，通过实体抽取算法抽取出句子中实体。

生成(实体对，文本)数据：根据抽取出的实体，结合该实体所属的文本，选择两个实体 E1，E2，对应的文本 S，从文本中生成(<E1，E2>，S)结构的数据。

(1)定义实体对关系。将实体对的关系定义为 6 个类别(见表 1)，按照 1 到 6 进行编号。

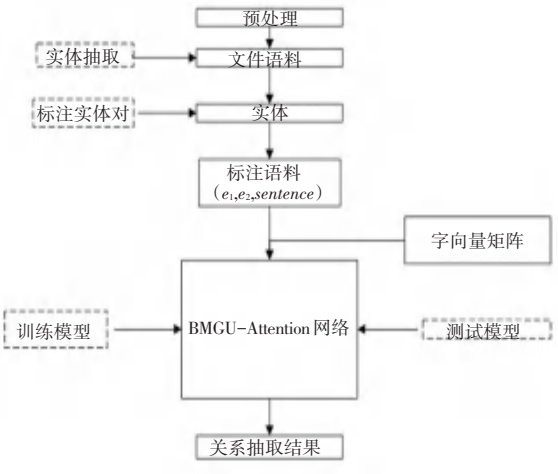

图 1　实体关系抽取框架

Fig. 1　Entity relationship extraction framework

表 1　关系定义

Tab. 1　Relationship definition

序号	对应关系名称	具体定义
1	位置	概念实体之间有如位于/发生在等关键词，表明实体间的位置关系
2	责任	一实体对另外一个实体负责，或者为领导关系
3	分类	一个实体为另一个实体的细分分类
4	发生	实体之间动词一般为发生
5	占有	一个实体对另一实体具有使用、安装等关系
6	其它	其它不再上述关系中的关系

(2)标注数据。根据煤矿领域训练集之间的基本关系，对要抽取的关系进行描述，对实体对-文本结构的数据进行标注，并且去除无意义的数据条目，优化训练数据。标注集格式为{<E1，E2>，R，S}，<E1，E2>表示实体对，R 为标注的关系，S 为包含实体对<E1，E2>的文本，一般为一个句子的长度。

(3)训练网络模型。

用字向量表示文本数据：

一个由 T 个字组成的句子 $S=(x_1,x_2,\dots,x_T)$，句子中的字 $x_i(1\leqslant i\leqslant T)$ 可以用一个向量 e_i 表示。Embedding 矩阵 $\boldsymbol{W}(\boldsymbol{W}\in R^{d\times|V|})$，词汇表 V 大小 $|V|$ 和矩阵 $\boldsymbol{W}$ 的维度相同，句子中的每一个字 x_i 可以用对应的向量 v(大小为 $|V|$)来表示，向量 v 是一个 one-hot 向量，只有在位置 i 处的值为 1，其余位置的值为 0。故向量 e_i 可以通过如下公式得到：

$$e_i=\boldsymbol{W}\boldsymbol{v}_i(1\leqslant i\leqslant T)$$

句子 $S=(x_1,x_2,\dots,x_T)$ 则可以表示为 $\{e_1,e_2,\dots,e_T\}$。

例如：句子 S 为“救援工人表示，由于井下的水

量大,这 2 台排水泵并不能满足要求,所以正在铺设其它水泵、水管",将每个字表示为 x_i,句子 S 可以表示为 $S=(x_1,x_2,\ldots,x_T)$,与 ***Embedding*** 矩阵相乘后句子 S 可以表示为 $S=\{e_1,e_2,\ldots,e_T\}$。

最后在神经网络层之后加入字级别的 Attention 机制,最终得到分类结果。

2.2 字向量+BMGU-Att 模型结构

字向量+BMGU-Att 模型由输入层、编码层、注意力层、输出层组成。注意力层由编码层得到的数据,根据对不同位置的权重计算得到输出,输出层的分类器得到最终的输出。模型结构如图 2 所示。

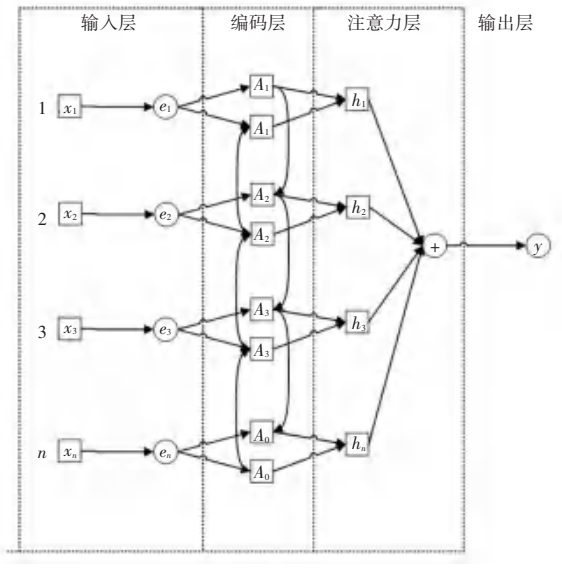

图 2 字向量+BMGU-Att 模型的结构示意图

Fig. 2 Schematic diagram of word vector +BMGU-Att model

2.2.1 输入层

输入层采用词嵌入机制对输入的文本数据进行降维处理,用字嵌入矩阵和对应位置的字向量相乘得到该字的字向量,最终将经过标注的长度为 T,格式为(实体对,包含该实体对的文本)数据转化为一个由 T 个字向量描述的数据,传入编码层。

数据由实体对:<E1,E2>和长度为 T 的文本 S,以及实体对<E1,E2>在文本中的关系 R 组成,结构为:<E1,E2>,R,S。数据为中文文本数据,输入层将每一个中文字符转化为字向量输入训练网络。

2.2.2 编码层

编码层使用 BMGU 网络作为训练网络,学习输入数据,对数据降维编码。BMGU 网络是一个双向循环神经网络,其中的网络单元由 MGU 单元构成。

BMGU 模型是基于 BRNN(双向循环神经网络,Bidirectional RNN,如图 4)算法的改进算法,使用 MGU 单元替换了传统单元。BRNN 是基于 RNN(循环神经网络如图 3)的改进算法,普通的 RNN 结构中,隐藏单元中的信息传递只能从"前"传到"后",位于后面时间步(位置)的单元可以学习到前面单元的状态信息,但是前面的单元不能学习到后面单元的状态信息。改进后的 BRNN 模型,在隐藏单元中增加了从后向前传递信息的单元,同一个输入不仅输入到正向的 RNN 单元中,同时也输入到反向的 RNN 单元中,因此 BRNN 模型中,一个时间步(位置)的单元同时接收到来自前后单元的状态信息。实际应用中,对文本信息进行处理是常常需要考虑到上下文对当前位置的词或字的影响,因此,采用双向循环神经网络更符合实际任务的需求。

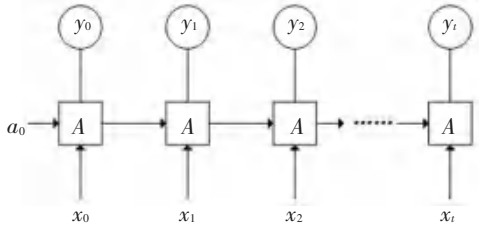

图 3 RNN 模型结构

Fig. 3 RNN model structure

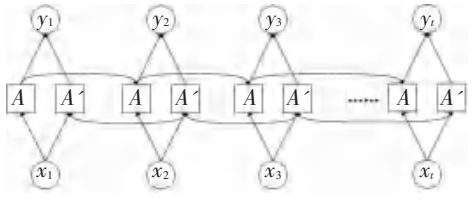

图 4 BRNN 模型结构

Fig. 4 BRNN model structure

MGU 算法由周志华等[11]人于 2016 年提出。MGU 算法是在 LSTM(Long-Short Time Memory)算法和 GRU(Gate Related Unit)算法的基础上得出。LSTM 算法中具有 forgate gate(遗忘门)、input gate(输入门)、output gate(输出门)三个门,是为了解决 RNN 算法在训练过程中发生梯度下降的问题而提出的。GRU 结构则是在 LSTM 的基础上进行简化,在解决梯度下降问题的同时也减少了隐藏单元中门的数量。GRU 中只有 2 个门:update gate(更新门)和 reset gate(重置门)。MGU 算法则在此基础上对门的结构进行再简化,只有一个门:forget gate(遗忘门),其单元结构如图 5 所示。

在 MGU 结构中,以 t 为当前时间步(位置),参数 x_t 代表在当前单元 6 的输入信息,h_{t-1} 代表上一时间步的 MGU 单元的状态信息,h_t 代表当前单元的状态信息,$\widetilde{h_t}$ 代表当前单元中的中间状态,f_t 表示通过 forget gate 得到的信息,b 代表不同位置的

bias。MGU 算法可描述如下：

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{1}$$

$$\tilde{h}_t = \tanh(W_h[f_t \cdot h_{t-1}, x_t] + b_h) \tag{2}$$

$$h_t = (1 - f_t) \cdot h_{t-1} + f_t \cdot \tilde{h}_t \tag{3}$$

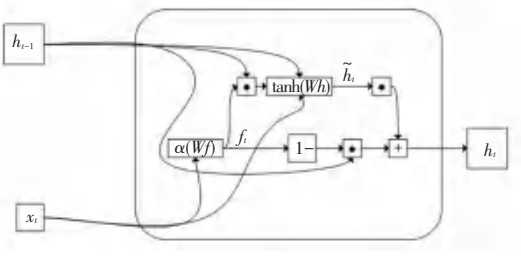

图 5　MGU 结构

Fig. 5　MGU structure

2.2.3　注意力层

注意力机制近来被广泛应用在问答系统、机器翻译、自然语言处理等领域中。本实验中，在双向 MGU 模型的输出后面使用注意力机制对编码结果进行优化。

双向 MGU 模型的输出结果为一个矩阵，设有矩阵 $\boldsymbol{w}$ 为对应权重矩阵，α 向量为各个字所应该具有的关注度，r 在对应的 α 下句子的表示，h 表示隐藏层的状态。具体计算公式如下：

$$M = \tanh(S) \tag{4}$$

$$\alpha = soft\ \max(w^T M) \tag{5}$$

$$r = S\alpha^T \tag{6}$$

$$h' = \tanh(r) \tag{7}$$

最终将 h' 作为输出层的输入。

2.2.4　输出层

输出层以隐藏层的状态作为输入数据，通过 soft max 算法对在句子 S 为条件的情况下分类的概率进行计算。定义 $\hat{y}$ 为最终要计算的预测结果，为一个维度为 d 的 one-hot 向量。计算方法如下：

$$\hat{y} = \arg\max(W_s h' + b_s) \tag{8}$$

3　实验结果分析

3.1　实验描述

本文使用的语料为煤矿安全网、煤矿事故网、安全管理网中有关煤矿管理和煤矿事故分析报告。经过预处理之后的实体对-文本数据共 2 M（总计约 20 000 条数据）。其中 1.2 M 作为训练语料，0.8 M 作为测试语料。实验中定义了 6 类关系，6 类关系描述见表 2、表 3。通过 Adam 方法控制学习率变化，最小学习率设定为0.000 5。

3.2　实验分析

训练过程中，使用相同的数据，设置相同的批次大小，相同的迭代次数，分别对以 LSTM、MGU、GRU 三种 RNN 单元为核心的模型进行训练，记录训练过程中的最高准确率和对应的 *loss* 值，具体信息见表 2。本文采用的方法和基于 LSTM、GRU 的方法相比，准确率基本一致，训练过程中的准确率相对接近，但在训练相同数量的数据时，使用 MGU 单元的训练时间更少，相比使用 LSTM 单元和 GRU 单元，本文采用的方法效率较高。

表 2　不同模型的训练过程中的最优结果

Tab. 2　Optimal results in the training process of different models

模型	*Accuracy*	*Loss*
BMGU-Att	0.96	2.184 14
BGRU-Att	0.96	4.992 28
BLSTM-Att	0.96	4.068 72

表 3　不同模型在相同数据量下训练消耗的时间对比

Tab. 3　Time comparison of training consumption of different models under the same amount of data

模型名称	消耗时间/min
BMGU-Att	13.2
BGRU-Att	20.15
BLSTM-Att	18.98

训练过程中不同模型的 *accuracy* 和 *loss* 值分别如图 6、图 7 所示。

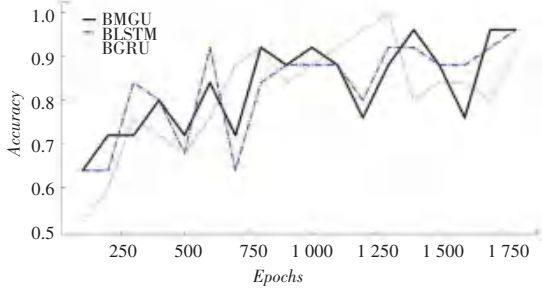

图 6　训练过程中 *Accuracy* 值变化曲线

Fig. 6　*Accuracy* value curve during training

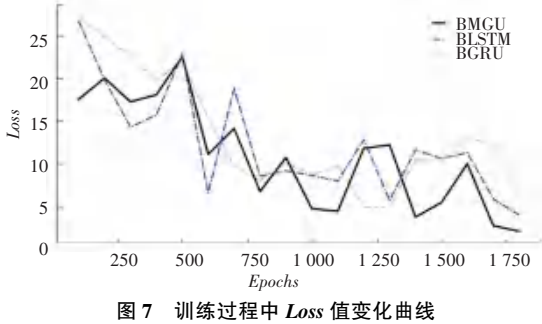

图 7　训练过程中 *Loss* 值变化曲线

Fig. 7　*Loss* value curve during training

根据上述图表中的数据,BMGU-Att、BLSTM-Att、BGRU-Att 三种模型的准确率相差不大,但 BMGU 模型的 *Loss* 值相对更小,对数据集拟合度更好。从不同模型占用资源的角度来看,可以明显地看到使用 MGU 作为 RNN 单元的 BMGU-Att 模型在训练过程中所消耗的时间显著低于另外 2 个模型的训练时间。因此,在相同情况下,BMGU-Att 模型的训练效率更高。使用验证集对已经训练好的 BMGU 模型进行验证,结果见表 4。

表 4 BMGU-Att 模型的验证结果

Tab. 4 Verification result of BMGU-Att model

类别	*Precision*	*Recall*	*F*1
1	0.736 842 1	0.666 666 7	0.7
2	0.811 074 9	0.778 125	0.794 258 4
3	0.842 857 1	0.769 565 2	0.804 545 5
4	0.882 352 9	0.784 883 7	0.830 769 2
5	0.752 873 6	0.850 649 4	0.798 780 5
6	0.718 535 5	0.779 156 3	0.747 619

由此可以看出,采用 BMGU 模型加注意力机制来进行分类,在总分类中能够达到较好的效果。在各类别的分类中,识别率会有所降低。不同类别的 *Precision*、*Recall*、*F*1 值差别较大,有可能是在训练数据中不同类别的训练数据所占比例不同所导致的。

4 结束语

本文采用 BMGU+Attention 模型,以字向量方法对文本输入进行处理,最终能够有效对实体对之间的关系进行抽取。通过实验,证明了该方法在煤矿领域中进行实体关系抽取是基本可行的。本文采用人工标注的实体对-文本数据训练模型,在文本数据中各个关系的数据比例不一,最终各个关系的分类准确率有一定差距,如何解决这两个问题是下一步研究的重点。

参考文献

[1] GUILLAUMIN M, MENSINK T, VERBEEK J, et al. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation[C]//2009 IEEE 12th International Conference on Computer Vision. Kyoto, Japan:IEEE, 2010:309-316.

[2] 邵堃, 杨春磊, 钱立宾,等. 基于模式匹配的结构化信息抽取[J]. 模式识别与人工智能, 2014, 27(8):758-768.

[3] CHEN Zhancheng, ZOU Bowei, ZHU Qiaoming, et al. Chinese negation and speculation detection with Conditional Random Fields [M]//ZHOU G, LI J, ZHAO D, et al. Natural Language Processing and Chinese Computing. Communications in Computer and Information Science. Berlin/ Heidelberg: Springer, 2013, 400:30-40.

[4] MANDLE A K, JAIN P, SHRIVASTAVA S K. Protein structure prediction using Support Vector Machine[J]. International Journal on Soft Computing, 2012, 3(1):67-78.

[5] JIVANI A G. The novel k Nearest Neighbor algorithm[C]// International Conference on Computer Communication and Informatics. Coimbatore, India:IEEE, 2013:1-4.

[6] KAMBHATLA N. Combining lexical, syntactic and semantic features with maximum entropy models for extracting relations [C]// Proceedings of Association for Computational Linguistics. Barcelona, Spain:ACL,2004:178-181.

[7] 陈宇, 郑德权, 赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报, 2012, 23(10):2572-2585.

[8] AIZERMAN M, BRAVERMAN E, ROZONOER L. Theoretical foundations of the potential function method in pattern recognition learning[J]. Automation & Remote Control, 1964, 25(6):821-837.

[9] 刘克彬, 李芳, 刘磊,等. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展, 2007, 44(8):1406-1411.

[10] ZHOU Peng, SHI Wei, TIAN Jun, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]// Meeting of the Association for Computational Linguistics. Berlin, Germany:ACL,2016:207-212.

[11] ZHOU Guobing, WU Jjianxin, ZHANG Chenlin, et al. Minimal gated unit for recurrent neural networks[J]. International Journal of Automation & Computing, 2016, 13(3):226-234.

(上接第 113 页)

[7] COUTO F M, SILVA M J, COUTINHO P M. Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors [C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany: ACM, 2005:343-344.

[8] LIN Dekang. An information-theoretic definition of similarity [C]//Proeeedings of the 15th International Conference on Machine Leaming.Madison,WI:dblp,1998:296-304.

[9] SCHLICKER A, DOMINGUES F S, RAHNENFÜHRER J, et al. A new measure for functional similarity of gene products based on Gene Ontology[J]. BMC bioinformatics,2006, 7(1): 302.

[10] LI Bo, WANG J Z, FELTUS F A,et al. Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins [J]. arXiv preprint arXiv:1001.0958,2010.

[11] WANG Dong, WANG Juan, LI Ming, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases [J]. Bioinformatics, 2010, 26 (13): 1644-1650.

[12] 李荣,曹顺良,李园园,等. 基于语义路径覆盖的 Gene Ontology 术语间语义相似性度量方法[J]. 自然科学进展,2006,16(7): 916-920.

[13] 李杰,初砚硕,程亮,等. 基于疾病本体的疾病相似性计算方法[J]. 生物化学与生物物理进展,2015,42(2): 115-122.