基于语义相似度计算的临床诊断自动编码算法研究

宁温馨 于 明

（清华大学工业工程系 北京 100084）

〔摘要〕 提出一种为中文临床诊断自动进行 ICD－10 编码的算法，利用分布式语义相似度计算方法计算文本语义相似度，考虑到中文的语言特点，不仅基于词语构建词向量，还基于汉字构建词向量，测试二者对查准率和查全率的影响。结果显示该算法在测试集上获得较高的准确率。

〔关键词〕 自动编码；语义相似度；分布式语义；ICD－10

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.02.011

Algorithmic Research on Automatic Coding of Clinical Diagnoses Based on Semantic Similarity Calculation *NING Wen－xin, YU Ming, Health Care Services Research Center, Department of Industrial Engineering, Tsinghua University, Beijing 100084, China*

〔**Abstract**〕 The paper proposes an algorithm which can implement ICD－10 coding automatically for clinical diagnoses in Chinese and calculate the semantic similarity of texts by the calculation method of distributed semantic similarity. In consideration to the linguistic features of Chinese, it constructs term vectors based on both words and Chinese characters and tests their influences on the precision ratio and recall ration. The results indicate that this algorithm has a higher precision ration in the test set.

〔**Keywords**〕 Automated code assignment; Semantic similarity; Distributional semantics; ICD－10

1 引言

电子病历（Electronic Medical Record，EMR）中包含了丰富的临床信息，这些信息有着很高的实际应用价值，如病人的健康状况跟踪、疾病的流行性分析、医疗服务质量控制、医疗决策支持等[1]。但是这些信息很难被直接利用，大都以文本的形式记录和存储，不利于后期的检索和总结归类，不便于做统计分析和决策支持[2]。医疗领域通用的解决方法是将这些医疗文本映射到一个标准的术语或编码系统中，如国际疾病分类系统（International Classification of Diseases，ICD）。现在疾病编码已被广泛应用于特定疾病发病率、死亡率的统计分析、医疗报销、病案管理以及卫生服务研究等领域[3]。在医疗机构中，编码工作由专门的编码员完成。编码员需要充分运用编码规则、医学术语等方面的大量知识，人工编码往往比较耗时且效率低下[4]。考虑到国内医院信息系统的普及程度越来越高[5－6]，医疗文本的积累也越来越快[7]，因此一个可以自动为医疗文本进行编码的计算机系统无疑将大大提升编码工作的效率，对于病案管理、医疗服务等领域有着很大意义。鉴于此提出一种对临床诊断自动分配 ICD－10 编码的算法，疾病和诊断的编码是医学编码中的重要部分，而 ICD－10 则是我国疾病编码的

〔修回日期〕 2015－10－29

〔作者简介〕 宁温馨，在读博士研究生；通讯作者，于明 。

标准体系。国内学术界对于 ICD－10 在理论与应用层面均有一定程度的研究。

2 算法设计

2.1 概述

针对医疗文本的自动编码问题，国际上有过一定的研究，但均以英文为主，基于中文文本编码者寥寥无几。中文词与词之间不存在分隔符，词语本身也缺乏明显的形态标记[8]，这些特征使得适用于英文的方法不适合直接应用于中文文本。本研究运用语义相似度理论，提出一套基于实例的自动编码算法。

2.2 基于实例的编码方法

在编码过程中，为一个给定的诊断分配 ICD－10 编码本质上相当于挑选出一个编码，使其能够尽可能准确、全面地描述该诊断文本所表达的信息。根据这一思想，提出了以下自动编码方法：给定一条诊断文本，挑选出一个 ICD－10 亚目的编码，使得这一编码对应的诊断实例与给定诊断最为相似。Pakhomov 等[9]的研究印证了这种基于实例的方法在临床诊断自动编码上的功效。但是他们的研究以梅奥诊所（Mayo Clinic）的诊断编码历史记录作为实例，而本研究基于的实例为国家卫生计生委的标准诊断库（下文简称标准诊断库）示例，见表1。

表1 标准诊断库示例

标准诊断	编码
古典生物型霍乱	A00.000
埃尔托型霍乱	A00.100
霍乱	A00.900
伤寒	A01.000
伤寒性肝炎	A01.001
伤寒性脑膜炎	A01.002

标准诊断库是由世界卫生组织国际分类家族合作中心带头编制，由国家卫生计生委统计信息中心于2013 年向全国发布的。标准诊断库由一系列标准化的中文诊断及其对应的 6 位数编码构成。ICD－10 亚目编码 A01.0 在其中包含了 3 条诊断，分别为“伤寒”、“伤寒性肝炎”和“伤寒性脑膜炎”。每个 4 位的亚目编码在标准诊断库中平均包含 2.28 条诊断。标准诊断库为诊断书写和编码选择提供了一个很好的参考。基于标准诊断库中的诊断编码实例，按照如下的方式为一个给定的诊断分配编码：

$$code = \underset{c \in C}{\operatorname{argmax}} sim(D, T_c)$$

其中，D 为给定的诊断文本，T_c 由标准诊断库中所有编码包含于 c 的诊断连在一起得到，例如，$T_{A01.0} = T_{A01.000} + T_{A01.001} + T_{A01.002}$ =“伤寒，伤寒性肝炎，伤寒性脑膜炎”。从某种程度上说，T_c 可以视为编码 c 的文本化描述。此外，C 为 ICD－10 中所有亚目编码的集合，sim 为两段文本的语义相似度函数。文本分词采用中科院计算所研发的基于层叠隐马尔科夫模型的汉语词法分析系统 ICTCLAS[8]。

2.3 文本语义相似度计算

通过选取标准诊断库中对应的实例与给定诊断最为相似的编码作为最终编码结果，因而文本语义相似度计算对于算法的性能至关重要。Mihalcea 等[10]提出了一种基于词语相似度和词语特异性的文本相似度度量方法。给定两段文本 T_1 和 T_2，其相似度由如下方式确定：

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in \{T_1\}} (maxSim(w, T_2) \cdot idf(w))}{\sum_{w \in \{T_1\}} idf(w)} + \frac{\sum_{w \in \{T_2\}} (maxSim(w, T_1) \cdot idf(w))}{\sum_{w \in \{T_2\}} idf(w)}\right)$$

其中，$maxSim(w, T)$ 为词语 w 与文本 T 中词语的最高语义相似度；$idf(w)$ 为词语 w 的逆文本频率，即语料库中的文档总数除以包含词语 w 的文档数目，反映了词语的特异性，idf 值越高，特异性越强。这里之所以没有利用经典的 $tf－idf$ 加权，是因为文本 T 内并未消除重复词语，实际上考虑了文本内的词汇频率与 $tf－idf$ 的思想是一致的。这一度量方法在文本语义相似度计算中取得了很好的效果，因此将其应用于文本的自动编码算法中。词语的逆文本频率以合作医院的入院记录为语料库计算得到。

2.4 词语语义相似度计算

词语或术语间的语义相似度计算方法可以分为两类，一类运用现有知识库中的分类信息来表达词语的语义[11-13]，另一类从大量的文本中获得词语的分布式统计量来表示语义[14-16]。考虑到本文主要针对医学领域的词语（构成诊断的词语大都属于医学领域），而我国目前并没有像 UMLS 或 SNOMED - CT 一样完备且公开的医学术语知识库，同时一般的中文词语知识库对医学词汇的覆盖程度较低，因此采用第 2 类方法，即分布式语义的方法计算词语的语义相似度。分布式语义相似度算法所基于的一个共同假设为，如果词语的上下文语境是类似的，那么词语就是相似的[17]。Patwardhan 和 Pedersen[16] 由此提出了一种语境向量的方法，该方法在国外生物医学领域的语义相似度计算研究中是最常用的一个分布式方法。这一方法基于语料库构建出一个共词矩阵，矩阵中的每一行对应一个词向量，词向量中的每一维对应语料库中的一个词语，w 的词向量中的每个元素代表词语 w 与该维度对应的词语在语料库中一个定长窗口内共同出现的频数。这样，一个词语的词向量记录了它与哪些词语经常共同出现，反应了该词语的上下文语境，进而表达了其语义。因此，两个词语 w_1 和 w_2 的语义相似度可以用它们词向量的夹角余弦值来衡量：

$$sim(w_1, w_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1||\vec{v}_2|}$$

其中 $\vec{v}_1$ 和 $\vec{v}_2$ 分别是两个词语的词向量。采用这一方法来计算词语的语义相似度。但是由于这一方法是在英文背景下提出的，英文的词语之间存在而中文却没有明确的分隔符，所以还需要额外对中文语料库进行分词处理。由于分词工具无法保证100%的准确性，因此按照原方法不可避免地会构建出错误的词向量。针对这一现象提出了一种新方法，词向量的维度不再对应词语，而是对应汉字。即不再用词语而是用汉字来构建词向量，以表达分布式语义。这一想法来源于中英文的差异。英文中的字母（character）一般不表达任何语义，如“apple”（苹果）中，“a”或“p”对“apple”的语义均无贡献；而在中文中，汉字（Chinese character）却可以表达一定的语义，如“苹果”中，“果”就有“水果”之意，指出了“苹果”的一个属性。因此，本文除了基于词语构建词向量之外，还基于汉字构建词向量，以计算词语的语义相似度。

从合作医院收集 54 136 份入院记录作为构建词向量的语料库，对这些文档均进行了预处理，包括去除停用词、数字和标点。中文分词同样采用 ICTCLAS。在基于词语的方法中，窗口长度设为目标词周围 3 个词语；而在基于汉字的方法中，窗口长度设为目标词周围的 7 个汉字。

3 算法评价

为了评价算法的实际性能，从合作医院的电子病历中抽取了 16 330 条诊断编码历史记录作为测试集。每条记录包含一条诊断及其对应的 ICD 编码。使用查准率（Precision）、查全率（Recall）和 F 值衡量算法的性能。查准率和查全率定义如下：

$$Precision = \frac{\sum_c tp_c}{\sum_c tp_c + \sum_c fp_c}$$

$$Recall = \frac{\sum_c tp_c}{\sum_c tp_c + \sum_c fn_c}$$

其中 tp_c、fp_c 和 fn_c 分别代表编码 c 的真正类（正确编码的实例）、假正类（被算法错误编码的实例）和假负类（被算法忽略的正确实例）。F 值则综合考虑精准度与查全率，为二者的调和平均值：

$$F - score = \frac{2Precision \cdot Recall}{Precision + Recall}$$

4 实验结果

在我国疾病编码通常至少编至亚目一级，即 4 位编码，因此重点关注自动编码算法在亚目层面的准确性。算法在两种语义相似度计算方法（基于词语和基于汉字的语境向量法）下的实验结果，见表 2。可见，基于词语的分布式语义相似度计算方法比基于汉字的方法收获了更好的性能，在测试集上的自动编码 F 值高达 0.893 2，说明中文词语比中

文字符蕴含了更多的语义。但是基于汉字的方法在编码准确性上并没有过大的差异，同样收获了较高的 F 值（0.875 6）。说明基于汉字来表达分布式语义只会对准确性有轻微的影响，但却可以完全省去中文分词的工作，因而这一新的思路是可以接受的。

表2 算法实验结果

项目	精准度	查全率	F 值
语义（词语）	0.931 2	0.858 2	0.893 2
语义（汉字）	0.901 4	0.851 3	0.875 6
精确匹配	0.854 5	0.757 9	0.803 3

为了探究语义相似度在临床诊断自动编码的任务中是否不可或缺，提出一种精确匹配的方法做对比。在精确匹配中，两个词语完全相等时相似度为1，否则相似度为0。这种词语的匹配方法在英文中比较常见。通过表2 可以发现，相比于计算词语的语义相似度，精确匹配法下的编码 F 值大幅度地降低，仅为0.803 3。说明语义相似度在临床诊断自动分类与编码中是不可或缺的，同时也说明对英文的处理方法不一定都适用于中文。

5 算法应用设计

由于自动编码算法很难能达到100% 的准确率，因此在实际应用中，该算法还应在一定程度上具备检测错误输出结果的能力。为此引入置信度这一指标。当对给定诊断 D 输出编码 c^* 时，该输出的置信度为 $sim(D, T_{c^*})$。置信度可以反映出输出的编码能够多大程度地描述给定的诊断，进而反映出编码结果的可靠性。图1 显示了测试集中当输出结果置信度大于给定阈值时的准确率及这部分结果所占的百分比。可见阈值设得越高，输出结果置信度大于阈值的比例越低，而这部分结果的准确性也越高。当阈值设为0.875 时，测试集中81.45% 的诊断能以高于该阈值的置信度编码，这部分诊断的编码准确率为0.971 1（F 值）。这一准确率水平已经达到了很高的程度，因而这部分编码结果可以不经人工检验直接输出。剩余18.55% 的诊断编码 F 值为0.571 4，所以人工检验必不可少。

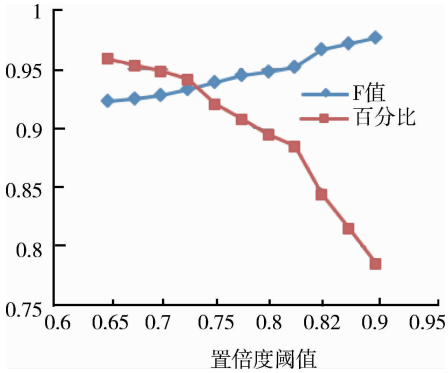

图1 置信度大于给定阈值时的编码准确率及百分比

由此制定了自动编码算法在实际应用中的流程：输入诊断运行算法后得到编码结果和置信度，当置信度大于0.875 时直接输出编码，否则将编码结果交由编码员人工检验或重新编码。依照这一流程，80% 以上的诊断能够由算法以高准确率（0.971 1 的 F 值）自动编码，只有不到20% 的诊断需要提交人工检验，从而大大降低了编码员的工作量并能够显著提升编码工作的效率。

6 结语

本研究提出了一种针对中文临床诊断进行自动编码的算法，综合利用现有的资源，将语义相似度计算理论应用于医学领域，同时在实现过程中考虑了中文的特点。算法在测试集中获得了0.893 2 的 F 值。经过实际应用的设计，超过80% 的诊断能够被算法以高准确率自动编码。该算法能够大大降低编码员的工作量并显著提升我国疾病编码工作的效率。

参考文献

1 Hornberger J. Electronic Health Records: a guide for clinicians and administrators [J]. JAMA, 2009, 301 (1): 110 - 110.

2 Meystre S M, Savova G K, Kipper - Schuler K C, et al. Extracting Information from Textual Documents in the Electronic Health Record: a review of recent research [J]. Yearbook of Medical Informatics, 2008, (35): 128 - 144.

3 O'Malley K J, Cook K F, Price M D, et al. Measuring Diagnoses: ICD code accuracy [J]. Health Services Research, 2005, 40: 1620 – 1639.

4 Pereira S, Névéol A, Massari P, et al. Construction of a Semi – automated ICD – 10 Coding Help System to Optimize Medical and Economic Coding [C]. MIE. 2006: 845 – 850.

5 凌红, 陈龙. 医院信息系统发展案例分析 [J]. 医学信息学杂志, 2013, (12): 16 – 20.

6 贾末, 王永刚, 沈韬, 等. 医院信息系统性能优化策略探讨 [J]. 医学信息学杂志, 2014, (35): 28 – 31.

7 苏韶生, 杨勇, 何远源, 等. 电子病历文档管理系统设计与关键问题实现 [J]. 医学信息学杂志, 2015, (1): 23 – 27.

8 刘群, 张华平, 俞鸿魁, 等. 基于层叠隐马模型的汉语词法分析 [J]. 计算机研究与发展, 2004, 41 (8): 1421 – 1429.

9 Pakhomov S V S, Buntrock J D, Chute C G. Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example – based and machine learning techniques [J]. Journal of the American Medical Informatics Association, 2006, 13 (5): 516 – 525.

10 Mihalcea R, Corley C, Strapparava C. Corpus – based and Knowledge – based measures of text semantic similarity [C]. In: Proceedings of the 21st National Conference on Artificial Intelligence. 2006, 6: 775 – 780.

11 Rada R, Mili H, Bicknell E, et al. Development and Application of a Metric on Semantic Nets [J]. Systems, Man and Cybernetics, IEEE Transactions on, 1989, 19 (1): 17 – 30.

12 Wu Z, Palmer M. Verbs Semantics and Lexical Selection [C]. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. 1994: 133 – 138.

13 Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy [C]. In: Proceedings of the 14th International Joint Conference on Artificial intelligence. 1995: 448 – 453.

14 Lund K, Burgess C. Producing High – dimensional Semantic Spaces from Lexical Co – occurrence [J]. Behavior Research Methods, Instruments, & Computers, 1996, 28 (2): 203 – 208.

15 Landauer T K, Dumais S T. A Solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge [J]. Psychological Review, 1997, 104 (2): 211.

16 Patwardhan S, Pedersen T. Using WordNet – based Context Vectors to Estimate the Semantic Relatedness of Concepts [C]. In: EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together. 2006, 1501: 1 – 8.

17 Harris Z. Distributional Structure [M]. New York: Oxford University Press, 1985: 26 – 47.