基于词性标注的中医症候名语料库

文/游正洋　王亚强　舒红平

摘要

文章对中医症候名语料库进行研究分析，并建立一个中医症候名的中英文对齐语料库。该语料库可以帮助识别中医医疗记录中易混淆的症候名。同时设计了一种标注方法对症候名数据集进行标注。语料库能够对中医临床症候名研究提供帮助。

【关键词】自然语言处理　文本挖掘　语料库　中医症候名　词性标注

1 前言

在西方国家，中医是一种与西医互补的、可替代的医学系统；但是在亚洲国家，中医在几千年前就已经被用来治疗各种疾病。中医是研究人体生理学、病理学和预防治疗人类疾病的一门学科。目前的中医理论基于宇宙原理和中国哲学，包含了整体论、分化、阴阳和五行理论。中医的医疗方法专注于提高人体自我控制系统和人体内部环境的协调来增强人体对疾病的抵抗力。中医的治疗相对复杂，其医学思想与现代西方医学有很大不同，因此，在大部分研究者看来中医的研究难度相对较大。目前，文本挖掘越来越多地被应用在中医临床记录地研究中，而自然语言处理方法被考虑作为一种工具来提高文本挖掘在中医临床记录研究中的潜力。文本挖掘的基本目标是找出文本中的潜在内容和萃取潜在知识，如内在联系、简洁的用户模式等。伴随者中医可用数据的迅速增加，迫切的需要浏览这些从大量文献中获取的资源数据。中医症候名就是其中最重要的数据之一。

然而，中文，尤其是中医症候名，具有非常丰富的语义。在不同的上下文中，一个中文汉字可能含有超过一种语义；不同的中文汉字的组合又会带来另外的含义。在中医症候名中，一些症候名的含义相同但是症候名称不一致。例如，“肾气虚证”和“肾气亏虚证”的症候名称不一致，但是含义是相同的。由于症候名称的不一致导致医生之间的交流效率受到影响。因此，提高中医临床记录中的症候名识别程度变得很有必要。

	syndrome	of	deficient	cold	in	uterus	null
胞							1
宫						1	
虚			1				
寒				1			
证	1						

图 1：中英文对齐标注矩阵

图 2：中英文对齐语料库示例

根据以上需求，研究建立了中医症候名的中英文双语对齐语料库。语料库提供了一种通过英文识别有混淆语义的中医临床症候名的方法。该语料库同时也可供中医文献和临床记录的文本挖掘使用。

2 语料库建立方法

2.1 数据处理

中医症候名原始数据从病人的诊断记录和治疗记录中获取。从中共获取了 812 个未处理的症候名称。为了使数据集更简洁，我们将原始症候名进行切分。在原始症候名中包含小括号和中括号两种类型的词汇。小括号中的汉字表示可以忽略；中括号中的汉字表示可以进行替换。例如，“心气 (亏) 虚证”表示“亏”可以忽略，则可将“心气 (亏) 虚证”拆分成“心气虚证”和“心气亏虚证”两个词；“冲任失 [不] 调证”表示“不”可以被替换，则可将“冲任失 [不] 调证”拆分为“冲任失调证”和“冲任不调证”。经过处理，数据集一共包含了 1129 个症候名。接下来我们分别使用人工翻译和机器翻译将每个症候名翻译为英文，以便对翻译质量进行对比。我们将一份翻译标准作为对比依据。将处理后的症候名再分割为单个汉字,；例如,“心气亏虚证”被分割为“心”、“气”、“亏”、“虚”、“证”5 个汉字。

2.2 标注方法

语料库使用矩阵来对 1129 个中医症候名进行标注。矩阵的第一列为被分割的中医症候名汉字，第一行为该症候名的英文翻译。如果拆分的汉字与英文单词对应，则标记为“1”；如果在英文翻译的单词中没有与汉字对应则在“null”列标记“1”。图 1 为中英文对齐标注矩阵。

我们建立了 2 个中英文对齐数据集用于对比参考。一个数据集通过翻译工具进行翻译，另一个通过人工进行翻译。我们对两者的翻译质量进行实验对比。

3 实验

3.1 翻译质量

我们使用一份翻译标准分别对工具翻译和人工翻译进行对比评估。使用翻译工具对症候名进行翻译后与翻译标准进行对比，在 1129 个症候名中工具翻译与翻译标准相同的词为 6 个，不同的有 1123 个。使用人工翻译对症候名进行翻译则有 1124 个词与翻译标准相同，5 个词与翻译标准不同。从统计数据可以看出，工具翻译与翻译标准对比差距较大，而人工翻译则与翻译标准差异较小。说明人工翻译比工具翻译更为准确。为了说明语料库的信度和翻译难度，我们引入了 kappa 系数对数据进行分析。

3.2 数据集的kappa系数分析

Kappa 系数是一种广泛使用的评估者之间

表 1：工具翻译的 kappa 值统计参数

		工具翻译 1		
		相同	不同	症候名总数
工具翻译 2	相同	4	2	6
	不同	4	1119	1123
	症候名总数	8	1121	1129

表 2：人工翻译的 kappa 值统计参数

		人工翻译 1		
		相同	不同	症候名总数
人工翻译 2	相同	744	5	749
	不同	380	0	380
	症候名总数	1124	5	1129

的评分一致性的指标。Kappa 系数公式为：

$$K = \frac{P_0 - P_e}{1 - P_e} \quad (1)$$

其中，p_0 为实际一致率，p_e 为随机一致率。如果一致率完全相同则 K=1。K 值计算结果为 -1 到 1 之间，其绝对值越小说明一致性越低。

为了分别计算工具翻译与人工翻译的 kappa 值，我们分别建立了两组工具翻译和人工翻译的数据集。两组工具翻译采用不同的翻译工具，两组人工翻译同样使用不同的翻译人员进行翻译。表 1 展示了工具翻译的 kappa 系数矩阵。从矩阵可计算出工具翻译的 kappa 值为 0.583，说明不同的工具翻译具有中等的一致性。表 2 展示了人工翻译的 kappa 系数矩阵，其 kappa 值为 -0.009，说明不同的人工翻译之间具有较低的一致性。

4 语料库相关分析

中医症候名的中英文双语对齐语料库共有 1129 个症候名、5618 个分割汉字和 4591 个英文翻译。在语料库中，我们使每个汉字都与一个英文翻译对齐，如图 2 所示。我们通过中英文的映射标记了中英症候名之间的联系。语料库提供了中英文的症候名对齐，该语料库可以用于具有混淆语义的中医临床症候名的识别，同时也可用于中医文本挖掘的研究。

5 结论

中医症候名的中英文双语对齐语料库完成了 3 个相关任务：症候名预处理，翻译和症候名分割，症候名标注与对齐。该语料库可作为中医症候名研究的基础，同时可以帮助研究者更有效和更精确地识别临床中医症候名。语料库也存在以下不足：语料库数据集数量偏小。在今后的研究中会不断的增加新的中医症候名，使识别准确率更加精确。

参考文献

[1]Fang Y,Huang H,Chen H.TCMGeneDIT:a database for associated traditional Chinese medicine, gene and disease information using text mining[J]. BMC Complementary and Alternative Medicine,2008.

[2]Wang S,Li Y,Devinsky O,et al. Traditional chinese medicine[J]. Complementary and alternative therapies for epilepsy,2005:177-182.

[3]Lu A P,Jia H W,Xiao C,et al.Theory of traditional Chinese medicine and therapeutic method of diseases[J]. World journal of gastroenterology:WJG,2004,10(13):1854.

[4]Hafner C.Introduction to Traditional Chinese Medicine (Out of Print)[J]. 2006.

[5]Ananiadou S,Kell DB,Tsujii [J].Text mining and its potential applications in systems biology. Trends Biotechnol.2006(24):571-579.

[6]Feng Y,Wu Z,Zhou X,et al.Knowledge discovery in traditional Chinese medicine:state of the art and perspectives[J]. Artificial Intelligence in Medicine,2006,38(03):219-236.

[7]Viera A J,Garrett J M.Understanding interobserver agreement:the kappa statistic[J].Fam Med,2005,37(05):360-363.

[8]Cohen J.A coefficient of agreement for nominal scales[J].Educational and psychological measurement,1960,20(01):37-46.

作者单位

成都信息工程大学软件工程学院　四川省成都市　610225