医疗领域文本结构化

丁祥武，张夕华+

（东华大学 计算机科学与技术学院，上海 201620）

摘 要：现有通用分词工具对医疗专业术语的识别效果不理想，影响了医疗文本结构化的效果。针对该问题，提出一种基于词向量的新词发现方法，利用新词发现过程中构建的词库抽取信息，得到结构化数据。使用 Google 开源词向量工具 word2vec 训练文本，将词映射到抽象的 n 维向量空间；根据词与词之间的得分、词的左右信息熵和在文本中的词来发现新词，把发现的新词加入用户自定义词库；设计信息抽取规则，根据发现的关键词提取对应的关键信息，将其组织为结构化数据。实验结果表明，用该方法进行结构化处理在准确率上比传统方法提高了 10%，在效率上比传统方法提高了 18%。

关键词：医疗文本；中文分词；词向量；信息熵；信息抽取

中图法分类号：TP311 **文献标识号**：A **文章编号**：1000-7024（2017）10-2873-06

doi：10.16208/j.issn1000-7024.2017.10.050

Text structuralization in medical field

DING Xiang-wu, ZHANG Xi-hua+

(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: The effects of general-purpose word segmentation tools recognizing medical terminology are not ideal, which greatly affects the accuracy of text structure. In view of the above problem, a method of discovering new words based on word embedding was put forward. Google open source word vector tool word2vec was used to train text and to map the words into abstracted n-dimensional vector space. New words were found using the information entropy, word frequency and the internal associative strength between word and word. The key information was extracted according to the key words. As a result, the structured data were made of key words and key information. Experimental results on real medical data show that the accuracy of the proposed method is improved by 10% compared to traditional method and the efficiency of the proposed method is improved by 18% compared to traditional method.

Key words: medical text; Chinese word segmentation; word embedding; information entropy; information extraction

0 引 言

文本结构化处理一般要经过以下 4 个步骤：分词、构词分析和词典处理、句法分析、领域分析。通常使用如下 3 种方式进行中文分词：基于词典、基于统计、基于规则。文献［1］对基于词典的正向最大匹配算法进行描述，文献［2］讲述了逆向最大匹配算法，文献［3］主要介绍了全二分最大匹配算法。基于统计的互信息的概率统计算法见文献［4］，文献［5］主要描述了 N-Gram 算法，文献［6］主要描述了基于组合度的分词决策算法。这些方法都被用于现在流行的分词工具中，如中科院的 ICTCLAS[7]、复旦大学的 FNLP[8] 以及开源的轻量级 IK Analyzer[9] 等。医疗文本具有以下 3 个特点：①特定用语；②同义词表达；③缩略语。这 3 个特点使得目前通用的中文分词工具对医疗文本分词的效果不是特别地明显。例如，对专业术语例如“皮细胞”、“胶质物”无法做到正确分词。针对上述问题，本文使用 Google 开源词向量工具 word2vec[10] 将文本中的词转化为向量，根据词向量计算词与词之间的得分，得分的高低表示它们之间的内部结合度的大小，再利用内部结合度以及词的左右信息熵、词频等统计信息，发现新词，并构建用户自定义的词库，将构建的词库添加到分词工具中，完成词库的扩展，对文本重新分词。最后，设计信息

收稿日期：2016-08-24；**修订日期**：2017-08-16

基金项目：上海市科技行动计划基金项目（15511106900）；上海市智慧城市建设基金项目（2015 年 1 月至 2016 年 12 月）；上海市信息化发展资金基金项目（XX-XXFZ-05-16-0139）

作者简介：丁祥武（1963-），男，上海人，博士，副教授，研究方向为大数据与列存储技术、分布式处理、多核与众核并行技术等；+通讯作者：张夕华（1992-），男，江苏淮安人，硕士研究生，研究方向为大数据处理技术、分布式处理等。E-mail：1244914259@qq.com

抽取规则，根据关键词库(指标名称)获取关键信息(指标值)，并将它们构造为结构化数据。

1 基本定义

令 $S=\{W_1, W_2, \cdots, W_i, \cdots, W_n\}$，$S$ 代表文本中的某条记录，W_i 表示文本中该记录的第 i 个词。词 W_i 的长度 L 等于 S 中的词 W_i 包含单个字的个数。W_i 在文本中出现的次数记作词频 Cnt。词位 Loc 顾名思义是指词 W_i 在记录 S 中的位置。

定义 1 $E_{L(w_i)}$ 定义为 W_i 的左信息熵，$E_{L(w_i)}=-\frac{1}{n}\sum_{w\in A}Cnt(w,w_i)\log\frac{Cnt(w,w_i)}{n}$。$n$ 是文本中词 W_i 的词频；A 是一个词集合，包含了文本中词 W_i 左边的所有词。

$E_{R(w_i)}$ 定义为词 W_i 的右信息熵 $E_{R(w_i)}=-\frac{1}{n}\sum_{w\in B}Cnt(w_i,w)\log\frac{Cnt(w_i,w)}{n}$。$n$ 是文本中词 W_i 的词频；B 是一个词集合，包含了文本中词 W_i 右边的所有词。

定义 2 M 定义为词 W_j 与词 W_i 之间的互信息，$M=\log\frac{p(w_iw_j)}{p(w_i)p(w_j)}$。

定义 3 新词(*neword*)定义为一个词集合，$neword=\{w \mid Cnt(w)>t_1 \wedge E_{L(w)}>t_2 \wedge E_{R(w)}>t_3\}$。$t_1$ 表示词频，t_2 表示左信息熵的阈值，t_3 表示右信息熵的阈值。

2 结构化处理流程

文本分析的前提是需要对文本进行结构化处理。文本结构化处理过程主要可以分为 3 个阶段：①预处理；②中文分词；③信息抽取。图 1 是对文本数据进行结构化处理的流程。

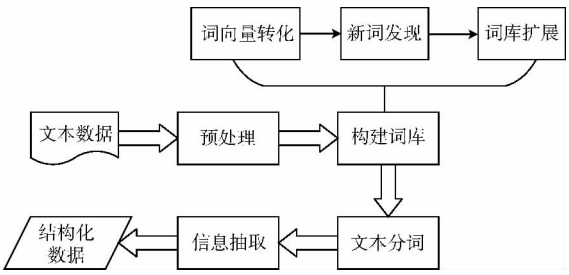

图 1 结构化处理的整体框架

预处理主要负责清洗文本中的数据，使得文本中不存在重复信息，没有明显的表达性错误，并且保证数据具有一致性。中文分词主要是对预处理后的文本进行分词，通用的分词工具无法正确地识别专业术语。通用的分词工具都有各自的词库，无法正确识别专业术语的原因是自带词库中不包含某个领域的专业术语，所以需要对词库进行扩展，添加自定义词库用于中文分词。从图 1 可以看出，构建词库需要将文本中的词转换为词向量，然后通过词向量来发现新词，最终将新词扩展成词库。

词向量转化是将文本中的词转化为向量，通过对向量进行处理，可以将词的操作转化为对向量的操作。CBOW 和 Skip-gram[11,12] 是 word2vec 的两种训练模型，它们的共同点是去除了非线性隐层，缩短训练时间。新词发现主要是查找本文中所述的专业术语，专业术语没有被包含在分词工具自带的分词词库中。构建一个用户自定义词库，用于存放新词发现过程中发现的新词。然后把用户自定义词库添加到分词工具中，医疗文本中的专业术语能够被分词工具正确地识别。最后根据专业术语和特定的抽取规则进行信息抽取，得到结构化数据。

3 构建词库

由于现有的分词工具对专业术语的分词效果并不理想，需要构建一个专业术语词库用于文本分词。本文主要使用 word2vec 将文本中的词转化为词向量，然后利用词向量来发现新词，构建用户自定义词库，即专业术语词库。将用户自定义词库添加到分词工具中，分词工具对专业术语能够正确地分词。

3.1 词向量

文中提到的词向量都是通过 word2vec 生成，word2vec 把训练文本中的词映射到 N 维实数向量。向量之间不是毫无联系，而是代表词与词之间的潜在语义关系。

本文在生成词向量时使用的训练模型是 Skip-gram，它将隐含层去掉，提高了训练的效率。尽管神经网络的隐含层很重要，一般不会去掉隐含层，但是实践证明了去除隐含层的可行性。从图 2 可以看出，Skip-gram 模型的基础是预测概率 $P(W_i \mid W_j)$。假设存在一个词组序列，把它表示为 $W_1, W_2, \cdots, W_j, \cdots, W_t$，Skip-gram 模型的目标是最大化 $\frac{1}{t}\sum_{1\leqslant t\leqslant T}\sum_{-c\leqslant j\leqslant c, j\neq 0}\log p(w_{t+j} \mid w_t)$ 的值。本文中使用的训练窗口大小是 5，即每个词在预测概率时只考虑前 5 个词和后 5 个词。

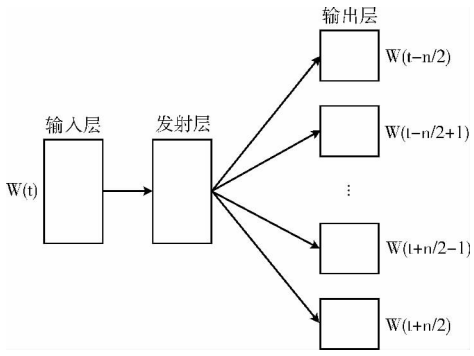

图 2 Skip-gram 模型

3.2 新词发现

新词发现的作用是识别文本中的“新词”，“新词”是

特指某个领域的专业术语，文中研究的是医疗领域。本文基于词向量提出一个新词发现算法，根据每个词的得分、左右信息熵和词频，发现医疗文本中的新词（专业术语）。具体的算法流程如图 3 所示。

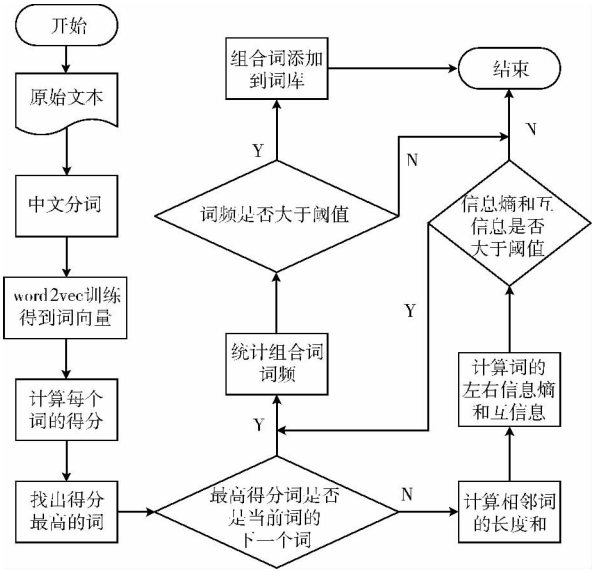

图 3　算法流程

从图 3 可以看出，发现新词主要分为 3 个步骤：①得到词向量；②计算每个词的得分。第 i 个词 W_i 的得分 $Stag_i$ 是一个集合，包括词 W_i 与所有词的内部结合度得分，$Stag_i=\{Stag_{i,1}, Stag_{i,2}, \cdots, Stag_{i,j}, \cdots, Stag_{i,n}\}$，$Stag_{i,j}$ 表示文本中第 i 个词 W_i 和文本中第 j 个词 W_j 之间内部结合度的得分，是 $\overrightarrow{N_i}$ 的单位向量与 $\overrightarrow{N_j}$ 的单位向量的内积，即 $Stag_{i,j}=\overrightarrow{Avg_i}\cdot\overrightarrow{Avg_j}$，$\overrightarrow{Avg_i}$ 是 $\overrightarrow{N_i}$ 的单位向量，即 $\overrightarrow{Avg_i}=\frac{\overrightarrow{N_i}}{\overrightarrow{N_i}\cdot\overrightarrow{N_i}}$。$Stag_{i,j}$ 可以反映出词 W_i 和词 W_j 之间的紧密程度，即内部结合度。令 Score 为词 W_i 的最高内部结合度，$Score=Max\{Stag_i\}$。找出与词 W_i 之间内部结合度得分最高的词 W_j，判断词 W_j 在文本中是否为词 W_i 的下一个词。如果是，则统计词 W_i 和词 W_j 构成的组合词 W_iW_j 的词频。如果 W_j 不是 W_i 的下一个词，需要计算词 W_i、词 W_{i+1}、词 W_{i+2} 的长度和。如果长度和不超过 5，把词 W_i、词 W_{i+1}、词 W_{i+2} 的长度和定义为 $Size$。再分别计算词 W_i 的右信息熵 $E_{R(w_i)}$、词 W_{i+1} 的左信息熵 $E_{L(w_{i+1})}$、词 W_{i+1} 的右信息熵 $E_{R(w_{i+1})}$ 和词 W_{i+2} 的左信息熵 $E_{L(w_{i+2})}$。如果长度和超过 5，则把词 W_i、词 W_{i+1} 的长度和定义为 $Size$。再分别计算词 W_i 的右信息熵 $E_{R(w_i)}$、词 W_{i+1} 的左信息熵 $E_{L(w_{i+1})}$。最后根据新词的判断条件（见定义 3），判断组合词是否为新词。具体的新词发现算法如算法 1 所示。

算法 1：GetScore

（1）计算词 W_i 的 $Score$。

（2）找出与词 W_k，该词与词 W_i 的内部结合度得分最高。

（3）比较 k 与 i+1 之间的大小关系。如果 i+1=k，构造一个新词 W，词 W 由词 W_k、词 W_i 拼接而成。然后，运行第（12）步。

（4）计算 3 个连续词的长度和 $Size$，分别是第 i，i+1 和 i+2 个词。如果 $Size>5$，跳转执行第（8）步。

（5）计算词 W_{i+2} 的左信息熵、词 W_{i+1} 的左信息熵、词 W_{i+1} 的右信息熵、词 W_i 的右信息熵。

（6）判断本文信息熵设置的阈值与第（5）步计算的 4 个信息熵的大小关系。如果阈值不是都小于 4 个信息熵的值，执行第（8）步。

（7）定义一个新词 W，词 W 由词 W_i、词 W_{i+1} 和词 W_{i+2} 组成，跳转执行第（12）步。

（8）判断设置的阈值是否小于词 W_{i+1} 的左信息熵以及词 W_i 的右信息熵。如果不是都小于，跳转第（15）步。

（9）计算互信息 M。

（10）比较 M 与设置阈值的大小关系。如果 M 小于设置的阈值，跳转执行第（15）步。

（11）定义一个新词 W，词 W 由词 W_i 和词 W_{i+1} 组成。

（12）计算词 W 的词频 Cnt。

（13）比较 Cnt 与设置阈值的大小关系。如果 Cnt 小于设置的阈值，跳转执行第（15）步。

（14）向用户自定义词库中添加 W。

（15）判断词 W_i 是否位于 S 的末尾。如果词 W_i 还有后续词，令 $i=i+1$，跳转到第（1）步。

4　信息抽取

创建一个用户自定义词库，该词库包含所有的“新词”。并把其添加到所用的分词工具中，利用用户自定义词库，对文本重新分词。然后对重新分词后的文本进行信息抽取，最终生成结构化数据。具体的信息抽取算法如 4.1 节和 4.2 节所述。

4.1　基于词库的信息抽取

本次的实验样本选取的是甲状腺穿刺数据，这类数据格式相对固定，描述简单。通过简单的人为统计，可以获取到指标值词库。有了指标值词库以及前面得到的用户自定义词库（指标名称词库），按照一定的规则，可以完成信息抽取，最终生成结构化数据。具体的算法如算法 2 所示。

算法 2：DirectoryStructuralization

输入：甲状腺穿刺文本中的任意一行数据

输出：该行数据对应的结构化数据

（1）对文本中的长句进行切分，生成多个短句。

（2）对多个短句一一进行中文分词操作。

（3）得到指标名称库 A，根据 GetScore 算法。

（4）得到指标值库 B，通过对样本数据统计分析。

(5) 遍历任一短句中的每一个词。如果 A 库或 B 库包含词 W_i，记录词 W_i 在短句中的位置，即词位 Loc。

(6) 对 A 库和 B 库中的词排序，排序的规则是按照词位排序，并将结果保存在一个新的数组中。

(7) 对新数组按照 B 库中的词进行划分，并把划分的结果存入到新的数组。

(8) 将数组中的第 2i+1 和第 2i 位置的数据，组织为结构化数据，其中 i 是自然数。

例如，样本数据中有这样一行数据："涂片见少量皮细胞和胶质物，提示结节性甲状腺肿"。按照上文的算法，可以得到 A、B 两个词库。A={"皮细胞"，"胶质物"}，B={"少量"}。对 A 和 B 两个词库按规则排序，可以得到一个新的数组 a[]={"少量"，"皮细胞"，"胶质物"}。然后再按照规则对数组 a 进行划分，得到的一个数组 b[]={"少量"，{"皮细胞"，"胶质物"}}。根据第(8) 步的操作，最终可以获得结构化数据{(皮细胞：少量)，(胶质物：少量)}。

4.2 基于词性分析的信息抽取

某些杂乱无章的文本，无法通过人为的统计得到准确的指标值词库(关键信息词库)。基于此类原因，本文提出了另一种信息抽取的方法：基于词性分析的信息抽取。该方法需要对词的词性进行标注，本文选用 ICTCLAS 分词工具，该分词工具具有词性标注的功能。具体的信息抽取算法描述如下。

算法 3：POSStructuralization

输入：甲状腺穿刺文本中的任意一行数据

输出：该行数据对应的结构化数据

(1) 得到指标名称库(关键信息库) A。通过使用上文所说的 GetScore 算法。

(2) 对短句进行词性标注。

(3) 判断短句中是否包含动词。如果不包含，跳转第(5) 步。

(4) 获取切分短句中的名词，如果该词出现在 A 库中，那么则跳转第(7) 步。

(5) 获取短句中的名词，并且该名词出现在 A 库中。

(6) 获取短句中的数词。

(7) 把得到的名词和数词生成为结构化数据。

假设有一条记录："涂片见少量胶质物和皮细胞，提示结节性甲状腺肿"。首先进行短句切分，然后对切分后的短句中的词进行词性标注，可得到如下结果：①"涂片/n 见/v 少量/m 胶质物/n 和/cc 皮细胞/n"；②"提示/v 结节性/n 甲状腺/n 肿/v"。通过 POSStructuralization 算法，可得结构化数据{(胶质物：少量)，(皮细胞：少量)}。根据表 1 中的数据可以清晰地看出，上文中所说的两种结构化处理方法都能得出正确的结构化结果。

表 1 非结构化数据转化为结构化数据

原始记录	结构化数据
涂片见少量皮细胞和多量淋巴细胞，提示桥本甲状腺炎	皮细胞：少量 淋巴细胞：多量
涂片见少量皮细胞和胶质物，提示结节性甲状腺肿	胶质物：少量 皮细胞：少量
涂片见多量淋巴细胞，未见肿瘤细胞，请结合临床随访	淋巴细胞：多量 肿瘤细胞：未见

5 实 验

5.1 实验环境

本文中的分词是在 Windows 系统中进行，分词工具选用的是 ICTCLAS。词向量转换工具 word2vec 的版本选用的是 C++，生成词向量需要在 Linux 系统中对 word2vec 进行编译。词向量保存在本地文件，并且词向量是以二进制的形式呈现。MyEclipse 作为主要的程序开发平台，jdk 的版本是 1.7.0_09。计算机的操作系统为 Windows，内存 4 GB，硬件配置为 Intel(R) Core(TM) i5-2400CPU @ 3.10 GHz。

5.2 实验数据

本文实验的样本数据是甲状腺穿刺诊断数据，该数据来源于某家三甲医院。把数据从数据集中抽取并清洗，总共有 20 156 条记录。在分词时使用的停用词典是哈尔滨工业大学的停用词表，共 702 个停用词，主要用于去除部分没有价值的数据。默认的分词词典是 ICTCLAS 的核心词典，共收集 79 836 个词语，是目前比较规范的词典之一，广泛用于中文分词。

首先对原始数据随机抽样 1000 条，对这 1000 条样本数据进行人工的结构化处理，得到这 1000 条数据对应的结构化数据，得到的结构化数据作为一组基线结果 (Baseline)。然后根据 Baseline，验证本文方法的准确率、召回率和 F 值。

5.3 实验结果和分析

传统基于信息熵的算法[13]是在重复串查找的基础上，结合词内部模式的特征，依次判断互信息、邻接类别等统计量，对新词进行识别。没有设置词长度的阈值，是对所有的词进行判断。本文在此基础之上，添加了词向量和词长度阈值的设置，在新词发现的性能上得到了很大地提升。

本文做了两组对比实验：①词长度分析；②性能对比实验。

5.3.1 词长度分析

随机抽样医疗文本中的 1000 条数据，对样本数据中的词长度进行统计，来验证设置阈值的必要性。主要统计样本数据中已登录词和新词的长度，图 4 展示了已登录词和新词在 L(词长度) 上的各个取值。从图 4 可以清晰地看出，能够组成新词的词，这些词的长度大多是 1 或 2。如果

某个词长度超过 5，则该词不可能作为新词的一部分。

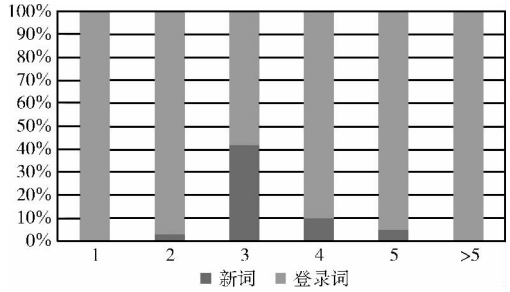

图 4　两类词在 L 上各个取值的比例

图 5 主要展示了新词与登录词在 L（词长度）上取值的分布。从图 5 可以得出一个结论，大部分词的长度都小于 5。因此，本文将词长度 L 的阈值设置为 5。如果一个词的长度超过 5，可以判断这个词不是一个新词。

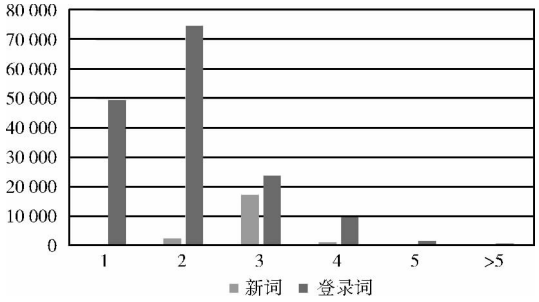

图 5　两类词在 L 上各个取值的分布

5.3.2　性能对比实验

对传统发现新词方法与本文所提方法在准确度方面进行分析，主要使用准确率、召回率以及 F 值进行准确性比较。这 3 个指标的计算公式如下

$$\text{准确率} = \frac{\text{正确发现新词个数}}{\text{发现新词总个数}} \times 100\%$$

$$\text{召回率} = \frac{\text{正确发现新词个数}}{\text{新词总个数}} \times 100\%$$

$$F = \frac{\text{准确率} \times \text{召回率} \times 2}{\text{准确率} + \text{召回率}} \times 100\%$$

对随机抽样的 1000 条医疗文本数据进行新词发现操作，采用的方法是本文所提的方法和传统新词发现方法。从图 6 得到 2 个结论：①准确率比传统新词发现算法提高 10%；②召回率和 F 值分别比传统新词发现算法提高了 9%。

通过简单的分析可以得出，本文新词发现算法效率高的原因是：相比传统的新词发现算法，省去了计算信息熵的步骤。如果能够满足条件，还需要统计该词在文本中的词频。如果词频很大，也满足设置的阈值条件，则认为是新词。

本次实验选用了 20 000 条甲状腺穿刺文本数据作为样本，分别对该样本数据进行新词发现，新词发现的算法分别为本文新词发现方法和传统的新词发现方法，最后对这

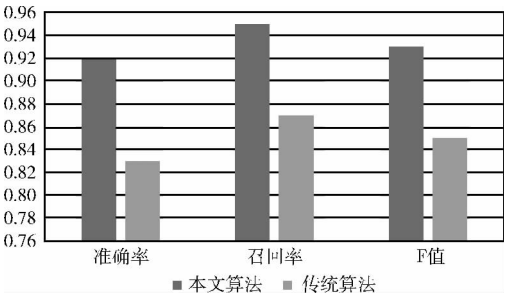

图 6　对比实验结果

两种方法的效率进行统计，统计的结果如图 7 所示。从图 7 可以看出，本文的新词发现算法比传统新词发现算法的效率大概提高了 18%。

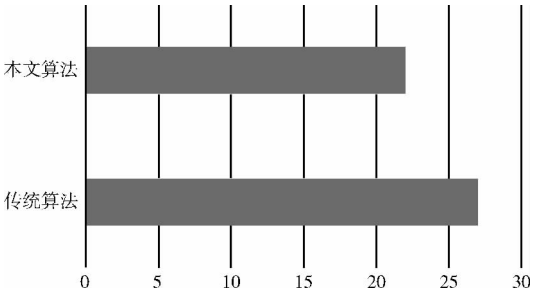

图 7　对比实验结果

6　结束语

本文提出了发现新词的方法 GetScore，该方法利用词向量和传统统计信息来发现新词，在没有使用语言规则作为特征的条件下，取得了较好的效果。再进行信息抽取，根据本文提出的信息抽取方法，可以将“无结构”的甲状腺穿刺文本数据转化为结构化数据。用该方法进行结构化处理在准确率上比传统方法提高了 10%，在效率上比传统方法提高了 18%。

未来的工作重心将转移到如何更好地利用词向量，建立模型，将深度学习应用在文本处理方面。

参考文献：

[1] WANG Ruilei, LUAN Jing, PAN Xiaohua, et al. An improved method of maximum matching algorithm of Chinese word segmentation [J]. Computer Application Software, 2011, 28 (3): 195-197 (in Chinese). [王瑞雷，栾静，潘晓花，等. 一种改进的中文分词正向最大匹配算法 [J]. 计算机应用与软件，2011，28 (3)：195-197.]

[2] DAI Pei, YANG Aimin. A reverse maximum matching based Chinese word segmentation algorithm [P]. CN: CN102999534A, 2013 (in Chinese). [代培，杨爱民. 一种基于逆向最大匹配的中文分词算法 [P]. CN: CN102999534A, 2013.]

[3] Wang S, Guo X, Mu X, et al. Bidirectional weight graph

transformation matching algorithm [C] //International Conference on Audio, 2015: 715-720.

[4] DU Liping, LI Xiaoge, YU Gen. Improvement of new words finding algorithm based on mutual information in the Chinese word segmentation system [J]. Journal of Beijing University (Natural Science Edition), 2016, 52 (1): 35-40 (in Chinese). [杜丽萍, 李晓戈, 于根. 基于互信息改进算法的新词发现对中文分词系统改进 [J]. 北京大学学报 (自然科学版), 2016, 52 (1): 35-40.]

[5] Behm A, Ji S, Li C, et al. Space-constrained gram-based indexing for efficient approximate string search [C] //IEEE International Conference on Data Engineering. IEEE Computer Society, 2009: 604-615.

[6] YUAN Dingrong, LI Xinyou, SHAO Yanzhen. Combination ambiguity resolution algorithm for Chinese word segmentation [J]. Computer Applications and Software, 2011, 28 (6): 57-58 (in Chinese). [袁鼎荣, 李新友, 邵延振. 用于中文分词的组合型歧义消解算法 [J]. 计算机应用与软件, 2011, 28 (6): 57-58.]

[7] Li X, Zhang C. Research on enhancing the effectiveness of the Chinese text automatic categorization based on ICTCLAS segmentation method [C] //4th IEEE International Conference on Software Engineering and Service Science, 2013: 267-270.

[8] Wen Bo, Li Hongguang. An approach to formulation of FNLP with complex piecewise linear membership functions [J]. Chinese Journal of Chemical Engineering, 2014, 22 (4): 411-417.

[9] Wang Z, Meng B. A comparison of approaches to Chinese word segmentation in Hadoop [C] //IEEE International Conference on Data Mining Workshop, 2014: 844-850.

[10] Mikolov T. Word2vec [EB/OL]. [2014-04-15]. http://code.google.com/p/word2vec/.

[11] Kenter T, Borisov A, Rijke MD. Siamese CBOW: Optimizing word embeddings for sentence representations [C] // Meeting of the Association for Computational Linguistics, 2016: 941-951.

[12] Shazeer N, Pelemans J, Chelba C. Skip-gram language modeling using sparse non-negative matrix probability estimation [J]. Computer Science, 2015.

[13] LIN Zifang, JIANG Xiufeng. The new word recognition based on the internal model of word [J]. Computer and Modernization, 2010 (11): 162-167 (in Chinese). [林自芳, 蒋秀凤. 基于词内部模式的新词识别 [J]. 计算机与现代化, 2010 (11): 162-167.]

(上接第 2868 页)

[2] Dai L, Zhang Y, Li Y, et al. MMW and THz images denoising based on adaptive CBM3D [C] //6th International Conference on Digital Image Processing. United States: SPIE, 2014.

[3] Yeom S, Lee DS, Son JY. Multi-level segmentation of passive millimeter wave images with Gaussian mixture modeling [C] // Terahertz Physics, Devices, and Systems V: Advance Applications in Industry and Defense. United States: SPIE, 2011.

[4] PAN Qilong. Passive millimeter wave detection and imaging system target detection and recognition algorithms [D]. Chengdu: University of Electronic Science and Technology of China, 2014 (in Chinese). [潘启龙. 无源毫米波探测成像系统目标检测与识别方法 [D]. 成都: 电子科技大学, 2014.]

[5] QIN Wenjie, ZHANG Guangfeng, LOU Guowei. Feature extraction method of PMMW radiation image based on security inspection [J]. Computer Engineering and Applications, 2012, 48 (28): 193-196 (in Chinese). [秦文杰, 张光锋, 娄国伟. 一种安全检查的无源毫米波图像特征提取方法 [J]. 计算机工程与应用, 2012, 48 (28): 193-196.]

[6] ZHANG Yan. Multi-target detection technology based on statistic and compressive sensing in SAR image [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2013 (in Chinese). [张焱. 基于统计特性和压缩感知技术的 SAR 图像多目标检测算法研究 [D]. 南京: 南京航空航天大学, 2013.]

[7] WEN Xin, HUANG Peikang, NIAN Feng, et al. Active millimeter-wave near-field cylindrical scanning three-dimensional imaging system [J]. Systems Engineering and Electronics, 2014, 36 (6): 1044-1049 (in Chinese). [温鑫, 黄培康, 年丰, 等. 主动式毫米波近距离圆柱扫描三维成像系统 [J]. 系统工程与电子技术, 2014, 36 (6): 1044-1049.]

[8] ZHAO Quan. CAFR detection of targets in SAR images based on statistical model [D]. Xi'an: Xidian University, 2012 (in Chinese). [赵全. 基于统计模型的 SAR 目标恒虚警检测方法研究 [D]. 西安: 西安电子科技大学, 2012.]

[9] XI Xianguo. Study on CFAR in radar clutter [D]. Dalian: Dalian Maritime University, 2013 (in Chinese). [席现国. 雷达杂波的恒虚警处理的研究 [D]. 大连: 大连海事大学, 2013.]

[10] Anastassopoulos V, Lampropoulos GA, Drosopoulos A, et al. High resolution radar clutter statistics [J]. IEEE Transactions on Aerospace & Electronics Systems, 1999, 35 (1): 43-60.

[11] DU Kun, WANG Wei, NIAN Feng, et al. Concealed objects detection in active millimeter-wave images [J]. Systems Engineering and Electronics, 2016, 38 (6): 1462-1469 (in Chinese). [杜琨, 王威, 年丰, 等. 主动毫米波图像的人体携带危险物检测研究 [J]. 系统工程与电子技术, 2016, 38 (6): 1462-1469.]