计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100087

基于多特征词语嵌入的日语文本聚类方法研究

于 娟[1] 李维婷[1] 曾心怡[1] 赵慧云[2]
1 福州大学经济与管理学院 福州 350108
2 中国移动通信集团福建有限公司福州分公司 福州 350108
(yujuan@fzu.edu.cn)

摘 要 针对传统日语文本词语表征的信息丢失及高维稀疏向量处理困难的问题,研究了日语文本词语提取和聚类方法。首先,根据日语的语言特性及改进的原子词步长法提取词语,并结合其统计特征、位置、词长和语义特征计算多特征融合权重值(Multi-attribute Fusion Weight,MFW),筛选得到文本特征词语集合,保留文本信息的同时实现特征降维;然后,以BERT加权特征词语MFW进行文本表示,并融合到以K-means++算法改进后的深度嵌入模型框架中,实现日语文本的聚类。在两个题材不同的日语文本数据集上进行实验,结果表明,该方法相较于已有方法在NMI和Purity指标值上均提升了5%以上,展现了良好的聚类效果。

关键词:日语文本挖掘;词语提取;文本表征;文本聚类;深度聚类
中图分类号 TP181

Japanese Text Clustering Based on Multi-attribute Word Embedding

YU Juan[1],LI Weiting[1],ZENG Xinyi[1] and ZHAO Huiyun[2]
1 School of Economics and Management,Fuzhou University,Fuzhou 350108,China
2 China Mobile Group Fujian Co.,Ltd. Fuzhou Branch,Fuzhou 350108,China

Abstract To address the problems of information loss in traditional Japanese text representation and the difficulty in processing high-dimensional sparse vectors,we study Japanese text word extraction and clustering methods. Firstly,the words are extracted using the improved atomic-word-step method based on Japanese linguistic characteristics. The Multi-attribute Fusion Weight (MFW) of the words is calculated combining their statistical features,positions,word lengths and semantic features so as to obtain a set of text feature words for retaining text information while reducing feature dimensionality. Then,Japanese texts are represented as the BERT-weighted MFWs of feature words,which is fused into the deep embedding model framework improved by the K-means++ algorithm to realize the clustering of Japanese texts. Experimental results on two Japanese text datasets with different topics show that the approach proposed in this paper improves both the NMI and Purity index values by more than 5% compared with the existing methods,which demonstrates a good clustering performance.

Keywords Japanese text mining,Word extraction,Text representation,Text clustering,Deep clustering

经济全球化和数字全球化的趋势使得多领域、多语言的文本数据迅猛增长,信息与经济社会发展的联系愈发紧密。处理和分析来自不同语言的文本信息,深度挖掘数据中隐含的知识和价值,帮助管理决策者更好地理解、组织和运用信息资源,推动全球经济文化等领域的交流与合作,已成为当前科学研究和实际应用中的迫切需求。其中,文本聚类是按内容将相似文本分组成簇的方法,旨在揭示文本数据中的潜在模式和结构,是文本数据挖掘的重要研究课题之一。目前,文本聚类被广泛应用于新闻信息聚合、用户行为分析、舆情分析等任务,为跨国组织的管理和决策提供着有力的信息支持。

日语是日本的官方语言,有一亿多以日语为母语的人,日语在世界范围内也有大量的学习者和使用者,是全球使用较为广泛的小语种之一[1]。同时,随着"一带一路"倡议的推进以及全球化进程的加快,中日两国的交流与合作不断加深,对日语文本信息分析和利用的需求也不断上升,对日语文本聚类研究愈发重要。由于日本在科技、文化等领域的影响,日语被广泛地作为第二外语学习,因此国内外的日语文本数据资源丰富。探索如何高效地挖掘和利用海量日语文本的潜在信息,开展日语文本聚类研究,对于推动中日跨文化交流、促进两国合作具有重要的研究价值。

相较于文本聚类方法研究较为成熟的中文和英文,日语文本具有其独有的特征。日语属于日本-琉球语系,具有黏着语的特点。与印欧语系不同,日语文本的词语之间连续排列,没有空格的间隔标识。与汉语不同,日语可以通过在词语前后黏着语素组成新的词以表达更丰富的语义(例如,在「茶」前黏着语素「お」构成词语「お茶」,表示更礼貌、尊敬的语义),具

基金项目:国家自然科学基金(72171090,71771054);福建省自然科学基金(2023J01393)
This work was supported by the National Natural Science Foundation of China(72171090,71771054) and Natural Science Foundation of Fujian Province(2023J01393).
通信作者:李维婷(weitingli1160@163.com)

有复杂的形态变化规则(例如,动词「書く」(译为"to write")根据时态和语态的不同可以变为「書いた」(译为"wrote")、「書かれた」(译为"was written")、「書いて」(译为"writing")、「書かない」(译为"do not write")、「書けば」(译为"were to write")等多种形态)。此外,日语文本中常常混杂多种不同类型的字符,例如:「デジタルトランスフォーメーション(DX)への投資」(译为"数字化转型(DX)投资")混合使用了汉字、片假名、平假名和英文字符。且日语文本中还存在许多文法混用的现象,例如,敬语「お願いします」和平語「頼む」常常交替使用。这给日语文本挖掘方法研究带来了不同的挑战。现有的日语文本聚类方法存在以下问题:在语种特点上,由于语言和语境差异,相对成熟的中文和英文文本聚类方法难以有效地直接迁移至日语文本聚类;在词语提取阶段,候选文本特征词语的提取主要基于固定词性搭配规则或 *N*-gram 模型,较少考虑词语的深层语义特征,特征词语的质量和完整性有待提高;在文本聚类阶段,传统文本聚类方法简单直观,但其存在难以提取文本语义信息、面对高维稀疏向量略显吃力等问题。

为了深化中日跨文化和跨领域的交流合作,推动日语文本聚类的研究与应用,针对上述问题,本文提出了一种基于多特征词语嵌入的日语文本聚类新方法。首先基于日语的语言特点,采用改进的原子词步长法[2]提取高质量的候选词语集合,同时计算词语的多特征融合权重值 MFW 以筛选特征词语集合,为下一步的文本分析任务提供更有价值的语义和主题信息;然后利用预训练模型 BERT(Bidirectional Encoder Representations from Transformers)[3]获取特征词语集合的嵌入表示,并结合 MFW 得到日语文档的向量表示,获得包含更多文本信息的文本表征;进一步结合以 K-means++算法改进的 DEC(Deep Embedding Clustering)[4]模型框架,同时学习文本表示和聚类分配,在改善文本表示的同时实现更优的聚类效果。

1 相关工作

本文聚焦于基于词语提取的日语文本聚类方法研究,为深入了解该领域的研究现状和进展,识别现存问题,下文从日语文本词语提取方法和日语文本聚类方法对本文的相关工作进行了梳理和总结。

1.1 日语文本词语提取方法

词语提取旨在从文本中提取有意义的特征词汇,为文本分析任务提供能够表达文本语义和主题的量化信息,减少计算复杂性的同时提高文本聚类的效果。文献[2]将文档中的词语分为原子词和合成词两类。原子词即为日语文本中的单词或是词素,其提取工作在预处理时的分词中就能得到较好地实现,如「電子」(译为"电子"),「情報」(译为"信息")等;合成词则是由原子词组合而成的具有完整意义的长词或短语,如「電子商取引」(译为"电子商务"),「情報システム」(译为"信息系统")等。

现有的文本词语提取方法研究主要包括基于规则的方法、基于统计的方法、基于图的方法、基于嵌入的方法和集成方法。基于规则的方法主要使用词汇、语法或其他语言结构等语言知识实现词语的自动提取。该方法可解释性强,且无需依赖大规模的语料支持,但规则的通用性有限,难以覆盖多种情况。基于统计的方法使用词语在语料库中的分布统计属性提取词语[5],主要包括 TF-IDF 算法、信息熵、互信息等。该方法简单高效,但缺点在于需要依赖足够大规模的语料库,且对于低频词的提取效果较差。基于图的方法常常被用于关键词语提取这一研究领域,其典型方法是 TextRank 算法。在此基础上,学者引入了更多的文本特征以构建词图中节点与边的权重,以实现算法的改进[6]。这一方法能够捕捉词语之间的复杂关系,但缺点在于,其受限于文本数据的质量和数量,准确率有限。基于嵌入的方法则通过嵌入向量来表示词语和文档的语义信息,并使用二者来衡量候选关键词与文档的语义相似性,以提取关键词语[7-8]。该方法考虑了词语的语义信息,但词语和文档序列长度的差异导致二者的文本表示并不充分匹配,影响了其在长文本词语提取的性能。集成方法则结合使用上述两种或两种以上方法实现文本词语提取[3,9]。它通过整合多种方法的优势,能够适应不同领域的任务需求,但其难点在于,针对不同任务需平衡不同方法的性能。

针对日语文本的词语提取方法研究较少,相关研究主要集中在日文分词、日文短语提取、日文关键词提取和日语命名实体识别等方面。日文分词常伴随着词性标注和词性还原,共同构成日文形态分析的过程。当前这一研究已较为成熟,能够实现较高的速度和准确率[10-11]。同时,基于日文分词研究,众多学者开发了多种日文形态分析工具,如 MeCab[12],spaCy[13],Jagger[11]等。在日文短语提取方面,文献[14]对文本数据进行分层形态分析,将其划分为词素,对词素的 *N*-gram 进行解读和计算,以提取出紧密相连的词素序列作为输出,使用统计方法对候选短语进行分类,并借助依存句法分析删除不恰当的短语。使用该方法从企业披露文件中提取战略短语的召回率较高,但准确率较低。在日文关键词提取方面,文献[15]利用固定词性搭配规则和 EmbedRank 算法提取候选词语,通过 MMR 算法对这些词语进行挖掘,以提取出更重要的关键词语,其提取误差为 5%~10%。在日语命名实体识别方面,文献[16]通过定义化学名称提取规则,从专利文本中提取并建立了化学名称语料库,进一步使用 Word2Vec 对词语集进行向量化,并使用机器学习方法对词语集进行二值分类,构建了一个化学物质名称提取模型,该模型的提取准确率可达到 80%以上。

1.2 日语文本聚类方法

文本聚类方法主要包括传统的文本聚类方法和基于深度学习的文本聚类方法。传统的文本聚类方法可分为基于划分的方法、基于层次的方法、基于密度的方法、基于模型的方法和集成方法,典型算法包括 K-means 算法、HAC 算法、DBSCAN 算法和 SOM 算法等。这类方法简单高效,应用广泛,但同时存在难以提取文本语义信息、面对高维稀疏向量略显吃力、对于大规模数据计算复杂度急剧上升等问题。随着深度学习的发展,许多研究开始探讨如何将深度学习与文本聚类相结合,并引入神经网络模型学习文本数据的高级抽象特征,基于深度学习的文本聚类方法应运而生。文献[4]提出了深度嵌入式聚类(DEC)模型,通过神经网络层将高维数据映射到低维空间,使用基于 KL 散度的聚类损失函数和嵌入损失函数进行训练,并在其中迭代优化聚类目标。文献[17]将 DEC 的深度嵌入聚类思想引入短文本聚类,并使用 Word2Vec 和平滑逆频率(SIF)嵌入进行向量表示,提出了

Self-Train 模型。文献[18]提出了一种集成序列信息和预训练文本编码器以提取深度语义特征的文本聚类框架 DFTC，同时该框架还结合了一个解释模块，以帮助理解聚类结果的意义和质量。文献[19]则将 Transformer 的句子向量表示与不同的聚类方法相结合，并将其成功应用于深度聚类任务。文献[20]在 DEC 模型的基础上引入对比学习，联合优化自上而下的聚类损失和自下而上的实例对比损失，提出一种基于对比学习的短文本聚类方法（SCCL）。文献[21]针对中长文本提出 DEC-transformer 模型，该模型引入预训练语言模型的迁移学习技术，并结合两阶段自反馈训练策略，在长文本聚类任务上取得了良好的效果。该类方法在文本聚类上取得了一定的成就，但仍然存在一些挑战，包括模型使用需要大量的计算资源、针对复杂长文本需要更加灵活和自适应的机制等。

当前日语文本聚类的研究则主要集中在聚类方法的应用上，关于聚类算法优化的研究较少；大部分研究更多使用传统的文本聚类方法，较少关注深度文本聚类模型。文献[22]使用 CBOW 模型、球形聚类算法以及 Siamese-LSTM 深度学习模型，开发并训练了一个能够计算日英跨语言财经新闻报道相似度的深度学习框架。文献[23]使用 BERT 和 HDBSCAN 聚类算法，设计了一个在搜索链接数据文本集中提取搜索动机关键词的聚类系统。文献[24]对 Word2Vec 获得的语义向量，使用 UMAP 非线性可视化和降维算法进行降维，应用 DBSCAN 聚类算法进行聚类，最后可视化提取客户物流不满意度。

综上所述，现有的词语提取和文本聚类方法大多侧重于中文和英文文本，而由于语言和语境差异，这些方法无法直接应用于日文文本上；针对日语文本词语提取和日语文本聚类方法的研究较少，且更多集中于日本国内。在日文词语提取上，当前候选词语的提取主要基于固定词性搭配规则或 N-gram 模型，候选词语的质量和完整性有待提高；候选词语的成词筛选主要基于词性特征和统计特征，考虑词语语义特征的研究较少，提取出的特征词语难以满足后续文本分析任务的要求，词语提取方法仍具有较大改进空间。在日文聚类上，当前的聚类方法更多关注传统的文本聚类方法，深度文本聚类模型的研究和应用则有待进一步深入，以更好地解决当前聚类方法所面临的挑战，实现更优的聚类效果。

2　基于多特征词语嵌入的日语文本聚类方法框架

基于文本特征词语嵌入的深度聚类思想，本文提出基于多特征融合的日语文本特征词语提取和基于多特征词语嵌入的日语深度文本聚类方法，方法框架如图 1 所示。在特征词语提取阶段，依据日语的语言特点优化文本预处理过程，改进原子词步长法以提取日语文本的候选词语集合，然后结合统计特征、位置、词长以及语义特征，计算候选词语的多特征融合权重值 MFW，以筛选后续文本分析所需的特征词语集合。在日语文本聚类阶段，使用 BERT 对筛选得到的特征词语集合进行编码，并加权 MFW 以构建日语文本集的文档向量表示，将文档嵌入向量输入自编码器进行预训练，进一步以 KL 散度损失联合训练自编码器和 K-means＋＋聚类算法，获得最终的聚类结果。

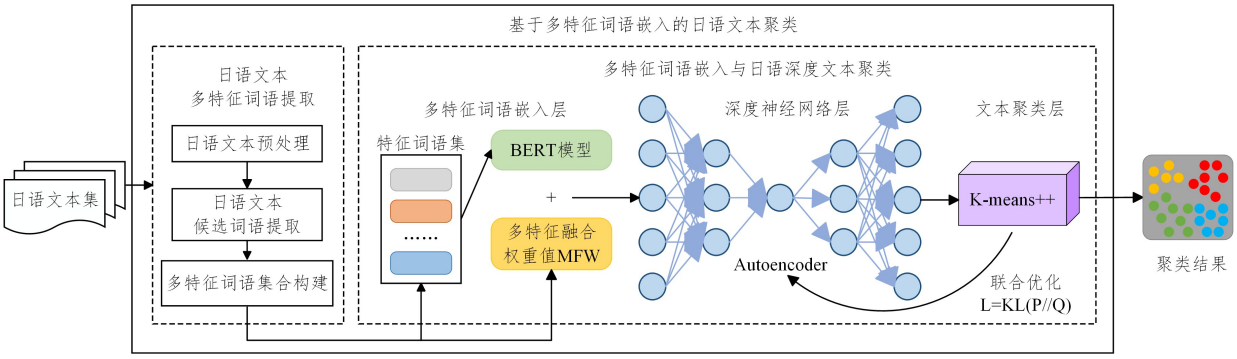


图 1　基于多特征词语嵌入的日语文本聚类方法流程

Fig. 1　Japanese text clustering based on multi-attribute word embedding

3　日语文本多特征词语提取

本文聚焦于多特征词语嵌入的日语文本聚类方法研究。其中，特征词语提取阶段主要包含日语文本预处理、候选词语提取和多特征词语集合构建三部分。

3.1　日语文本预处理

日语文本预处理主要包括文本清洗、形态分析（分词、词性标注和词性还原）、停用词去除 3 个步骤。首先，删除文本中的公式、符号和其他噪声元素。由于日语中汉字假名混合的现象，文本中往往使用全角和半角字符来表示不同的内容，因此有必要将其统一为全角字符的形态。在形态分析阶段，经过实验比较现成工具，选择分词质量和效率较高的 MeCab 工具，同时加入日本国立国语研究所开发的最新的 Unidic 词典[25]。然后，结合 spaCy 提供的停用词表，根据语料自身特点和多次的词语提取实验，总结出日语停用词表和停用词性表，如代名词、接续词和感动词等，用于过滤文本中的停用词，输出预处理后的日语文本。

3.2　日语文本候选词语提取

在日语文本候选词语提取阶段，改进了针对中文词语提取的原子词步长法[2]，将其迁移至日语文本处理任务上。相比于传统的候选词语提取方法，该方法既能满足提取文本任意 n 元词串的需求，保留更多的文本信息，又能够在保证提词质量的同时拥有高效的处理效率。其主要由频繁词串提取、子串删除和候选词串过滤三步组成。频繁词串提取，以原子词为步长，长度优先统计每个词串及其所有子串的频次，提取频次超过预设阈值的词串作为频繁词串，并经实验选取频次阈值为 2。子串删除，为避免父串截断而提取出不成词子串的问题，删除频繁词串集合中出现频次与父串相同的子串。候选词串过滤，使用不成词规则对输入的候选词串进行过滤，进而输出候选词语集合。结合日文特点和实验总结设置的不

成词规则如下。

1)词首尾过滤:修改首词词性为接尾辞的词串,删除接尾辞,保留剩余词串;同理,修改尾词词性为接头辞的词串。根据日语语言学规则,接尾辞通常附加在词干之后,接头辞附加在词干之前,针对这一特性和实验分析构建以上规则,例如,修改词串「型社会」为「社会」。

2)词长过滤:删除字符长度等于 1 或词串长度大于 10 的词串。字符长度等于 1 的词串大多为结构不完整、非日语实词且不具有实际意义的公共破碎子串,如「型」、「児」、「時」等,对于后续的文本分析任务来说意义不大;词串长度大于 10 的词串往往包含一些噪声信息,成词概率较低,因此考虑删除。

3)不成词词典过滤:过滤位于不成词词典中的词串。不成词词典是在多次词语提取实验的人工成词判定工作中、由日语专业人士判定为不成词的词串集合。

3.3 多特征词语集合构建

由于 TF-IDF(Term Frequency-inverse Document Frequency)简单直观,常用于评估特征词语对于文档的重要程度,但忽略了词语之间的语义信息,无法捕捉词语的上下文依赖关系,且对于长文档而言也无法体现位置信息。因此,本文尝试在 TF-IDF 的基础上,引入词语的位置特征、词长特征以及语义特征,构建词语多特征融合权重值 MFW,筛选能够代表一篇文档的特征词语集合。

一篇日语文档 D,经过第 3 章所述的日语文本预处理和日语文本候选词语提取操作,得到一个候选词语集 $W=\{w_1, w_2, \cdots, w_i, \cdots, w_q\}$,$w_i$ 表示候选词语集中的第 i 个词语,q 表示该候选词语集的词语数量。以词语 w_i 为例,具体说明本文构建文档 D 的特征词语的方法。

1)计算词语 w_i 的统计特征 $TF\text{-}IDF(w_i)$,如式(1)—式(3)所示。

$$TF(w_i)=\frac{n_{w_i}}{\Sigma_k n_{w_k}} \tag{1}$$

$$IDF(w_i)=\log\left(\frac{N_D}{N_{w_i}+1}\right) \tag{2}$$

$$TF\text{-}IDF(w_i)=TF \cdot IDF \tag{3}$$

其中,n_{w_i} 表示词语 w_i 在某一文档中出现的频次,$\Sigma_k n_{w_k}$ 表示该文档中所有词语的频数之和,N_D 表示语料库中的文档总数,N_{w_i} 表示词语 w_i 出现的文档数量之和。

2)计算词语 w_i 的位置特征 $Pos(w_i)$,如式(4)—式(6)所示。词语的出现位置通常对于判断词语的特征信息重要性具有很大的价值。对于一篇日语文章的创作而言,文首和文末常用于强调文章主旨、总结文章内容,出现在这些位置的词语往往更能传达文本的核心思想,更有可能成为文本表示的特征词语。由于特征词语具有词频大、在文本中分布较广的特性,难以捕捉特征词语在文章中出现的所有位置,因此本文参考文献[26],着重考量词语在文章中首次出现和末次出现的位置。

$$Position_{\text{first}}(w_i)=\frac{count\_before(w_i)+1}{\Sigma_k n_{w_k}-count\_before(w_i)} \tag{4}$$

$$Position_{\text{last}}(w_i)=\frac{count\_after(w_i)+1}{\Sigma_k n_{w_k}-count\_after(w_i)} \tag{5}$$

$$Pos(w_i)=N\left(\frac{1}{Position_{\text{first}}(w_i)+Position_{\text{last}}(w_i)}\right) \tag{6}$$

其中,$\Sigma_k n_{w_k}$ 表示一篇文档中所有词语的数量;$count\_before(w_i)$ 表示词语 w_i 第一次出现时,在其位置前面的所有词语数;$count\_after(w_i)$ 表示词语 w_i 最后一次出现时,在其位置后面的所有词语数;$N(\cdot)$ 为简单的最大最小归一化函数。

3)计算词语 w_i 的词长特征 $Len(w_i)$,如式(7)所示。词语的长度与其所蕴含的信息量存在一定的正相关关系。较长的词语往往携带更多的信息,能传达更具体、丰富的语义内容;相反,较短的词语表达的意义则更抽象、模糊,无法准确地反映一篇文档的主题。因此,本文引入词长特征,词长特征值越大的词语,其蕴含的主题信息量越大,成为文本特征词语的可能性越高。

$$Len(w_i)=\frac{len(w_i)}{\frac{1}{q}\sum_{i=1}^{q} len(w_i)} \tag{7}$$

其中,$len(w_i)$ 表示词语 w_i 的字符长度,$\frac{1}{q}\sum_{i=1}^{q} len(w_i)$ 表示一篇文档中词语的平均字符长度。

4)计算词语 w_i 的语义特征 $Sem(w_i)$,如式(8)所示。使用候选词语嵌入与文档块嵌入的相似度来筛选能够表示整个文本语义思想的特征词语。首先使用 BERT 分别提取候选词语和文档的向量表示,然后计算二者的余弦相似度值,用于衡量候选词语与文档的语义相似度。语义相似度越高的词语,其表示的主题或信息与文档的主要内容越相关,越能准确地反映一篇文档的主要特征和意义。本文比较选取了“bert-base-japanese-whole-word-masking”预训练模型作为文本表示的工具。相比传统 BERT 模型在字符粒度上进行分词,这一模型采用了全词掩码技术(Whole Word Masking),学习词语粒度的表征,从而提高模型对于日文词语的理解能力;此外,它使用 MeCab 日语分词工具对文本进行分词处理,并在大量的日语维基百科文章上进行了训练,这使其能够更加适应日语的语言特性,提高日语文本表示的准确性与效率。

$$Sem(w_i)=\frac{\boldsymbol{vec}_D \cdot \boldsymbol{vec}_i}{\| \boldsymbol{vec}_D \| \, \| \boldsymbol{vec}_i \|} \tag{8}$$

其中,$\boldsymbol{vec}_D$ 为文档 D 的向量表示,$\boldsymbol{vec}_i$ 为候选词语 w_i 的向量表示,$\boldsymbol{vec}_D \cdot \boldsymbol{vec}_i$ 表示二者的点积,$\| \boldsymbol{vec}_D \| \, \| \boldsymbol{vec}_i \|$ 表示两向量长度之积。

5)基于前述 4 步所得的特征值,计算词语 w_i 的多特征融合权重值 $MFW(w_i)$,如式(9)所示。其中,基于模型通用性和易用性的考虑,经实验,本文将 MFW 各部分权重参数设定为 1。

$$MFW(w_i)=TF\text{-}IDF \cdot (1+Pos(w_i)+Len(w_i)+Sem(w_i)) \tag{9}$$

以 MFW 降序排列候选词语,提取一定比例的词语输出为最后的特征词语集合。一方面,使用 MFW 值筛选特征词语能够兼顾词语各维度特征信息,使得到的特征词语集合能够更全面地反映文本内容,提高后续文本分析任务的效果;另一方面,提取一定比例的词语能够确保在处理不同长度的文本时,灵活地选择不同数量的特征词语,避免长文本择词少而损失文档信息,短文本择词多而包含无关信息的问题。

4 基于多特征词语嵌入的日语深度文本聚类

基于多特征词语集合,本文尝试以 K-means++算法改进 DEC 模型,提出一种基于多特征词语嵌入的日语深度文本

聚类方法(Japanese Deep Text Clustering Based on Multi-attribute Word Embedding,JDTC_MWE),主要由嵌入层、神经网络层和聚类层组成。

4.1　多特征词语嵌入层

嵌入层使用特征词语集合与 MFW 权重值加权以输出文档的向量表示。使用 BERT 模型对特征词语集合进行文本编码,得到词向量组,并结合特征词语的 MFW 权重值来表示一篇文档的文本向量。这一文本表示方法简单直观,能够在有效减少原始文本数据冗余的同时充分获取文本多维特征信息。文本向量计算如式(10)所示:

$$V(D)=\sum_{i=1}^{q}(vec_i \cdot MFW(w_i)) \tag{10}$$

4.2　深度神经网络层

神经网络层将文本嵌入表示输入自编码器进行预训练,对高维向量进行降维表示。自编码器(Autoencoder)是一种无监督学习的神经网络,由编码器(Encoder)和解码器(Decoder)两部分构成。它通过编码器将输入的数据压缩为低维表示,解码器则尝试对这一低维表示进行重构,并在迭代训练过程中使重构结果尽可能接近于原始输入数据。这一过程可表示为式(11)－式(13):

$$h=\mathrm{ReLU}(\boldsymbol{W}_e x+\boldsymbol{b}_e) \tag{11}$$

$$\hat{x}=\mathrm{ReLU}(\boldsymbol{W}_d h+\boldsymbol{b}_d) \tag{12}$$

$$\hat{x}\approx x \tag{13}$$

其中,x 为原始的输入数据,h 是编码器输出的、对原始数据编码后得到的低维隐藏表示,$\hat{x}$ 为解码器输出的、对低维隐藏表示进行重构的结果,ReLU 为各层的激活函数,$\boldsymbol{W}_e$ 和 $\boldsymbol{W}_d$ 为权重矩阵,$\boldsymbol{b}_e$ 和 $\boldsymbol{b}_d$ 为偏置向量。

4.3　文本聚类层

聚类层使用 K-means＋＋算法,并进一步将自编码器和 K-means＋＋聚类联合训练,通过不断迭代优化特征表示和聚类分配、降低聚类损失,以提升聚类的效果。首先,使用 K-means＋＋算法初始化聚类中心。该算法通过选择相互间距离尽可能远的点作为初始质心,为后续的任务提供一个相对可靠的初始聚类中心;然后,迭代进行以下 3 个步骤以更新网络权重与聚类中心,直到满足收敛条件。1)计算每个样本点的软聚类分配度。学生 t 分布(Student's t-distribution)在统计学中常用于小样本数据的统计推断,能够很好地处理数据的变异性和不确定性。软聚类分配度则基于学生 t 分布,计算每个样本点 z_i 与聚类中心 μ_j 的距离,并对其进行规范化,以此来估计样本点 i 到聚类簇 j 的概率。如式(14)所示,这里取自由度 α 为 1。

$$q_{ij}=\frac{(1+\|z_i-\mu_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}((1+\|z_i-\mu_{j'}\|^2/\alpha)^{-\frac{\alpha+1}{2}})} \tag{14}$$

2)计算基于软分配度的辅助分布。定义辅助目标分布 P 以迭代优化样本点概率分布 Q,旨在提高聚类结果的纯度和置信度,降低大簇对特征空间的影响,如式(15)所示:

$$p_{ij}=\frac{q_{ij}^2/\sum_{i'}q_{i'j}}{\sum_{j'}(q_{ij'}^2/\sum_{i'}q_{i'j'})} \tag{15}$$

3)在联合训练中最小化聚类损失。通过学习辅助分布 P 和概率分布 Q 的匹配程度,选取二者间的 KL 散度来度量聚类损失,聚类损失函数 L 表示为:

$$L=KL(P\|Q)=\sum_i\sum_j p_{ij}\cdot\log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{16}$$

5　实验分析

5.1　数据集与参数设置

本实验采用两个题材不同的文本数据集进行实验对比分析:日语 livedoor 新闻语料库(Livedoor News Corpus)[27] 和日语维基百科词条语料库(Japanese Wiki Corpus)[28]。日语 livedoor 新闻语料库是一个适用于日语文本分类的新闻数据集,本文从中随机选择 4 类共 1 350 篇文本构成数据集 1,平均每篇文本包含 465.36 个单词,总大小为 3.73 MB;日语维基百科词条语料库是一个基于京都词条的、开源的维基百科语料库,本文从中随机选择 8 类共 3 263 篇文本构成数据集 2,平均每篇文本包含 804.07 个单词,总大小为 25.6 MB。数据集内容如表 1 和表 2 所列。其中,本研究选择了中长文本作为实验数据集,两个数据集中的每篇文本的单词数均大于 300 个,其目的是为了后续研究特征词提取比例对于文本聚类效果的影响,保证能够在特征词语提取阶段尽可能获取更多的特征词语,保留更丰富的特征信息。

表 1　数据集 1 构成

Table 1　Composition of data set 1

类别	文本数/篇
IT 生活小窍门	335
家用电器	330
电影	335
体育	350
总计	1 350

表 2　数据集 2 构成

Table 2　Composition of data set 2

类别	文本数/篇
历史	666
文化	638
建筑	495
皇帝	471
学校	46
宗教	413
道路	191
地理	343
总计	3 263

本文的深度文本聚类框架参数设置为:在自编码器预训练阶段,使用 Adam 作为优化器,训练迭代次数设置为 100,训练批次大小设置为 256,隐藏层大小设置为 d:500:500:2 000:20,其中 d 为输入的嵌入向量的维度;在编码器和聚类层联合训练时,选取 SGD 优化算法,初始学习率设置为0.01,动量设置为 0.9,收敛阈值设置为 0.1%。

5.2　评价指标

基于本文数据集的构成情况,本实验从外部指标中选取常用的归一化互信息值(Normalized Mutual Information,NMI)和纯度(Purity)作为分析聚类结果的评价指标。假设文档总数为 N,将聚类后的文档集合表示为 $S=\{s_1,s_2,\cdots,s_u\}$,s_i 表示第 i 个聚类簇的集合,将已知文档类别的文档集合表示为 $Y=\{y_1,y_2,\cdots,y_v\}$,y_j 表示第 j 个类别的文档集合。

NMI 的计算式如式(17)、式(18)所示:

$$MI(S,Y)=\sum_{i=1}^{u}\sum_{j=1}^{v}P(s_i,y_j)\cdot\log\left(\frac{p(s_i,y_j)}{P(s_i)\cdot P(y_j)}\right) \tag{17}$$

$$NMI(S,Y)=\frac{2MI(S,Y)}{H(S)+H(Y)} \tag{18}$$

其中，$P(s_i)$表示某文档属于聚类簇 s_i 的概率，$P(s_i)=\frac{|s_i|}{N}$；$P(y_j)$表示某文档属于已知类别 y_j 的概率，$P(y_j)=\frac{|y_j|}{N}$；$P(s_i,y_j)$表示某文档既属于聚类簇 s_i 又属于已知类别 y_j 的概率，$P(s_i,y_j)=\frac{|s_i\bigcap y_j|}{N}$；$H(S)$表示聚类簇 s_i 的信息熵，$H(S)=-\sum_{i=1}^{u}P(i)\log P(i)$；$H(Y)$表示已知类别的信息熵，$H(Y)=-\sum_{j=1}^{v}P'(j)\log P'(j)$。

Purity 的计算式如式(19)所示：

$$Purity=\frac{1}{N}\sum_{i=1}^{u}\sum_{j=1}^{v}\max|s_i\bigcap y_j| \tag{19}$$

5.3 实验结果与分析

通过实验比较不同多特征加权方法、特征词语提取比例、文本表示方法以及文本聚类方法对文本聚类效果的影响，探究本文提出的 JDTC_MWE 方法的性能。同时，为避免偶然性，本节实验记录 10 次实验结果的平均值作为最终的聚类结果。

5.3.1 多特征加权方法的影响

为了探究本文提出的多特征融合权重值 MFW 对于文本聚类效果的影响，对 MFW 进行消融实验。MFW-S 表示在本文提出的多特征融合权重值的基础上去除语义特征，MFW-SL 表示在本文提出的 MFW 的基础上去除语义特征和词长特征，MFW-SLP 表示在本文提出的 MFW 的基础上去除语义特征、词长特征和位置特征。该实验使用简单的向量空间模型对特征词语集合的所有词语进行实验，分别使用 K-means 算法、K-means++算法及 HAC 算法进行实验对比，所得实验结果如表 3 所列。对表 3 分析可知：

1)MFW 特征加权方法的文本聚类效果优于其他减少特征信息的特征加权方法。这验证了多特征融合策略能够有效地捕捉文本的语义特征、词长特征、位置特征和统计特征，并整合多维特征信息，以提取能够全面反映文本内容和结构的特征词语，从而提升聚类的性能。

2)在两个数据集上，K-means++算法的表现均略优于传统的 K-means 算法。这一结果说明，K-means++算法通过优化初始质心的选择，可以有效提升聚类的质量和算法的稳定性。同时，在数据集 1 中，K-means++算法相对于 K-means 算法的提升不高。本文认为这是因为数据集 1 在向量空间上的点分布较为均匀，使得初始质心选择优化策略对最终聚类结果的影响较小，随机初始化也可能迅速收敛得到较好的解，因此二者的聚类性能并无较明显的差异。

3)对比两个数据集的聚类结果，数据集 1 的聚类结果显著优于数据集 2，这一差异可能主要归因于数据集规模和特征稀疏性两个方面。相比于数据集 1，数据集 2 包含的文档数量更多，其可能包含更多的噪声，增加了聚类任务的难度；同时，数据集 2 特征词语集合的平均数量大于数据集 1，这可能导致向量空间模型出现数据稀疏性问题，进而影响聚类的效果。

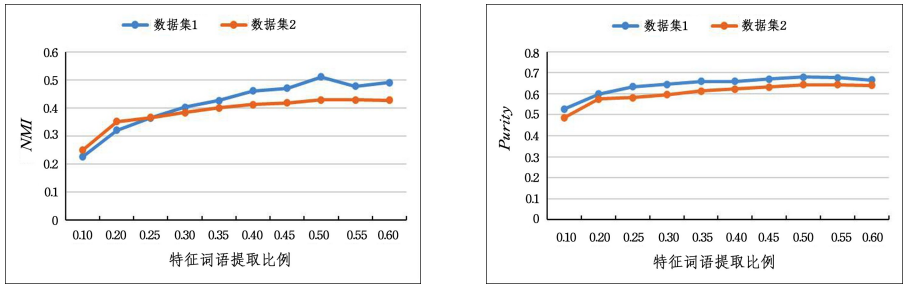
表 3 不同多特征加权方法的对比

Table 3 Comparison of different multi-attribute weighting methods

数据集	方法	K-means		K-means++		HAC	
		NMI	*Purity*	*NMI*	*Purity*	*NMI*	*Purity*
1	MFW	**0.3972**	**0.6640**	**0.3982**	**0.6653**	**0.3906**	**0.6456**
	MFW-S	0.3928	0.6646	0.3928	0.6646	0.3846	0.6581
	MFW-SL	0.3860	0.6559	0.3862	0.6563	0.3731	0.6253
	MFW-SLP	0.3748	0.6545	0.3751	0.6550	0.3610	0.6209
2	MFW	**0.1866**	**0.3759**	**0.1873**	**0.3778**	**0.1865**	**0.3812**
	MFW-S	0.1808	0.3677	0.1846	0.3686	0.1827	0.3783
	MFW-SL	0.1636	0.3622	0.1659	0.3653	0.1578	0.3604
	MFW-SLP	0.1601	0.3675	0.1645	0.3730	0.1533	0.3550

5.3.2 特征词语提取比例的影响

为了探究不同特征词语提取比例对文本聚类效果的影响，在候选词语集上按降序排列并提取了不同比例数量的特征词语集合，结合本文方法进行实验比较分析。提取比例数值为：0.1，0.2，0.2，0.3，0.35，0.4，0.45，0.5，0.55，0.6，共 10 组，在两个数据集上的实验结果如图 2 所示。



(a)NMI 指标下不同特征词语提取比例的对比　(b)Purity 指标下不同特征词语提取比例的对比

图 2 不同特征词提取比例的对比

Fig. 2 Comparison of different feature word extraction ratios

由图 2 分析可得：随着特征词语提取比例的增加，文本聚类的效果均呈现提升的趋势。这表明，较高的特征词语比例有助于捕获更为丰富的文本信息，进而提升文本聚类的质量。同时，观察两指标可以发现，这一值在提取比例为 0.5 左右达

到最大,后续提升比例值,指标值则呈现下降的趋势。这表明,继续提升特征词语比例可能引入过多冗余信息,影响聚类的效果。因此,后续选取特征词语比例为 0.2,0.3,0.4,0.5 这 4 组进一步开展探究实验。

其中,观察图 2(a)的实验结果可以发现,当特征词语提取比例在 0.1～0.25 时,数据集 2 的性能要高于数据集 1。这主要归因于两个数据集的不同特性。数据集 1 为新闻文本,其文本的类别边界相对模糊,如家电文本可能也包含 IT 生活文本的相关术语,因此提取较小比例的特征词语所呈现的聚类效果较差;而数据集 2 为维基百科文本,其文本聚焦于某一特定领域,类别区分度较为明显,少量的特征词语也能够使模型表现出较高的性能。

5.3.3　文本表示方法的影响

为了分析文本表示方法对文本聚类效果的影响,选取了文本表示阶段被广泛使用的 Word2Vec 平均法作为基线方法、结合本文多特征加权方法的 Word2Vec＋MFW 加权法和结合预训练语言模型的 BERT 平均法,与本文 BERT＋MFW 加权法进行比较。以上各文本表示方法均使用 DEC 算法进行文本聚类对比实验,在 2 个数据集和 4 组特征词语提取比例上的实验结果如表 4 所列。

表 4　不同文本表征方法的对比

Table 4　Comparison of different text representation methods

数据集	方法	*NMI*				*Purity*			
		0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
1	Word2Vec 平均法	0.0358	0.0450	0.0531	0.0580	0.2944	0.3036	0.3200	0.3329
	Word2Vec＋MFW 加权法	0.0483	0.0566	0.0596	0.0608	0.3585	0.3719	0.3156	0.3704
	BERT 平均法	0.2879	0.3729	0.4124	0.4529	0.5747	0.6156	0.6081	0.6341
	BERT＋MFW 加权法	**0.3175**	**0.3942**	**0.4512**	**0.4890**	**0.5956**	**0.6212**	**0.6444**	**0.6693**
2	Word2Vec 平均法	0.0949	0.1067	0.1197	0.1266	0.3164	0.3268	0.3324	0.3467
	Word2Vec＋MFW 加权法	0.1004	0.1119	0.1200	0.1367	0.3204	0.3288	0.3382	0.3547
	BERT 平均法	0.3047	0.3285	0.3703	0.3836	0.5318	0.5436	0.5847	0.5990
	BERT＋MFW 加权法	**0.3418**	**0.3728**	**0.3996**	**0.4231**	**0.5660**	**0.5897**	**0.6088**	**0.6315**

从表 4 的实验结果分析可得:

1)本文提出的 BERT＋MFW 加权法在文本聚类任务上展现出了较好的优势。一方面,相较于 Word2Vec 词向量表示,BERT 模型经过大规模语料预训练,使其能够学习到丰富的语言模式和知识,同时其双向 Transformer 结构,则使其能够捕捉到更加丰富的语义信息和上下文依赖,生成更加准确的文本表示;另一方面,相较于平均法,MFW 加权法通过为不同的特征词语分配不同的权重,突出了文本中的关键信息,降低了无关词语的干扰,使得文本表示更加聚焦于文本核心内容;二者共同作用,使得 BERT＋MFW 加权法在文本聚类任务中实现更好的效果。

2)随着特征词语比例的增加,不同文本表示方法的聚类效果也均呈现提高趋势。同时,即使调整特征词语比例,BERT＋MFW 加权法依然能保持相对稳定的优势,这表明该方法具有较好的稳定性和适应性,即无需依赖特征词语比例的调整也能获得较好的聚类效果。

3)对比两个数据集的聚类结果,数据集 1 的结果总体上仍优于数据集 2,但相较于使用向量空间模型的聚类结果,二者差距缩小。这也证明了 BERT 词嵌入在解决向量稀疏性问题上的优越性,体现了这一经大规模语料训练得到的高质量文本嵌入在提高文本聚类性能上的有效性。

5.3.4　文本聚类方法的影响

为了分析文本聚类方法对于文本聚类效果的影响,本文选取了文本聚类方法中简单直观且性能良好的 K-means 算法作为基线方法、优化后的 K-means＋＋算法和结合神经网络模型的 DEC 算法与本文提出的 JDTC_MWE 方法进行比较。该实验以上述实验中表现良好的 BERT＋MFW 加权法作为文本表示方法,实验结果如表 5 所列。

表 5　不同文本聚类方法的对比

Table 5　Comparison of different text clustering methods

数据集	方法	*NMI*				*Purity*			
		0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
1	K-means	0.2835	0.3637	0.4291	0.4450	0.5820	0.6668	0.7278	0.7313
	K-means＋＋	0.2838	0.3645	0.4293	0.4455	0.5822	**0.6672**	**0.7279**	**0.7317**
	DEC	0.3175	0.3942	0.4512	0.4890	0.5956	0.6212	0.6444	0.6693
	JDTC_MWE	**0.3201**	**0.4029**	**0.4606**	**0.5110**	**0.5987**	**0.6449**	**0.6593**	**0.6797**
2	K-means	0.2384	0.2963	0.3671	0.3946	0.4020	0.4773	0.5518	0.5817
	K-means＋＋	0.2391	0.2983	0.3696	0.3960	0.4026	0.4809	0.5545	0.5830
	DEC	0.3418	0.3728	0.3996	0.4231	0.5660	0.5897	0.6088	0.6315
	JDTC_MWE	**0.3513**	**0.3828**	**0.4118**	**0.4292**	**0.5746**	**0.5961**	**0.6226**	**0.6435**

由表 5 的实验结果可知:

1)在 NMI 指标上,DEC 算法均优于传统聚类方法,这一优势体现了深度聚类方法的优越性。DEC 算法通过使用神经网络构建的自编码器实现了对嵌入表示的特征提取和降维,同时通过自编码器和聚类的联合训练一体优化聚类和文本表示任务,有效地改善了文本聚类的质量。而在 Purity 指标上,DEC 在数据集 1 上的结果并没有比传统聚类方法更好。这可能是因为数据集 1 的规模较小,限制了神经网络模型性能的发挥。尽管如此,DEC 在 NMI 指标上的优异表现仍然表明其在处理复杂数据时的潜力。

2)本文提出的 JDTC_MWE 方法大体上优于其他聚类方法。这一结果表明,JDTC_MWE 方法不仅继承了 DEC 算法在文本聚类任务中的优越性,还通过将 K-means 替换为 K-means＋＋算法,解决了原始算法在初始聚类中心选取上的

不足，从而显著提升了文本聚类的性能。

5.3.5 聚类可视化

为了更加直观地展示聚类的结果，同时对聚类结果的实际意义进行深入探讨，本文选取了特征词语提取比例为 0.5 的数据集 1 文档矩阵，使用 t-SNE(t-Distributed Stochastic Neighbor Embedding)方法对本文模型的聚类结果进行可视化处理，不同颜色的点分别对应不同的类簇划分，如图 3 所示。在此基础上，进一步抽取各簇部分关键词，通过词云图进行可视化呈现，如图 4 所示。

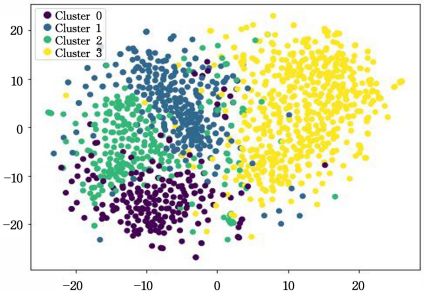


图 3 数据集 1 的聚类结果可视化

Fig. 3 Visualization of clustering results of dataset 1

(a)Cluster0 (b)Cluster1

(c)Cluster2 (d)Cluster3

图 4 聚类结果的词云图

Fig. 4 Word cloud map of clustering results

由图 3 可得，4 个类簇在聚类空间上大致呈现出分离状态，同一簇的文本自然聚集在一起；其中，类簇 0 和类簇 2 的点存在明显的重叠部分。由图 4 的词云图可得，类簇 0 中主要的关键词为「発売、フィルター、アンテナ、ケーブル」(译为“发售、过滤器、天线、电缆”)，将其归类于“家用电器”话题；而类簇 2 中包含有「カメラ、レンズ、ビデオ、防水、充電」(译为“相机、镜片、视频、防水、充电”)，表明其应与“IT 生活小窍门”话题相关。从这两个类簇所呈现的关键词来看，它们的内容在某种程度上具有相似性，因此在聚类结果可视化中出现类簇间重叠的现象。此外，类簇 1 中主要的关键词为「映画、前作、タイタン、魔物、主演」(译为“电影、前作、泰坦、魔物、主演”)，将其归类于“电影”话题；类簇 3 包含有「チーム、代表、試合、予選、優勝」(译为“团队、代表、比赛、预赛、冠军”)，则主要与“体育”话题有关。

结束语 为了推动日语文本聚类的深入研究与广泛应用，促进中日跨文化和跨领域的合作，本文提出了基于多特征词语嵌入的日语文本聚类方法。该方法充分利用日语文本特性，结合改进的原子词步长法实现高质量候选词语的提取，同时通过计算候选词语的多特征融合权重值(MFW)，以筛选得到特征词语集合；进一步结合使用特征词语集合、MFW 和预训练模型 BERT 获取文档的嵌入表示，并将这一嵌入层融合到以 K-means++算法改进后的 DEC 模型框架中，实现了基于多特征词语嵌入的日语文本聚类方法(JDTC_MWE)。

本文采用两个题材不同的日文数据集进行日语文本聚类方法的对比实验。实验结果表明，MFW 能够有效综合文本多维特征，提取较为令人满意的特征词语；结合特征词语集合的 BERT+MFW 加权法，结合 BERT 的深度语义理解和 MFW 的多特征加权策略，生成更加精确且富含信息的文本向量表示；最后，JDTC_MWE 方法通过结合 BERT+MFW 加权法的文本表示优势、深度嵌入聚类联合优化的策略以及 K-means++优化初始质心的改进，显著提升了文本聚类的效果，进一步展示了本文方法的优越性。

未来，将尝试把本文方法推广至其他语言，以增强其跨语言适用性和普遍性，使其能够在更广泛的场景中发挥作用；同时，考虑融合其他文本特征，如句法特征、情感特征等，以丰富特征词语的文本表示能力；此外，还将继续考虑优化深度文本聚类方法，通过添加其他神经网络模型或替换其他自编码器，进一步提升文本聚类的性能。

参 考 文 献

[1] Wikipedia. Japanese language[EB/OL]. [2024-01-31]. https://en.wikipedia.org/wiki/Japanese_language.

[2] YU J, DANG Y Z. Word extraction method combining part-of-speech analysis and string frequency statistics[J]. Systems Engineering Theory and Practice, 2010, 30(1): 105-111.

[3] JACOB D, MING-WEI C, KENTON L, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[4] XIE J, GIRSHICK R, FARHADI A. Unsupervised Deep Embedding for Clustering Analysis [C]// Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR, 2016: 478-487.

[5] DING X Y, WANG L C. Research on Optimized Calculation Method for Weight of Terms in BBS Text[J]. Information studies: Theory & Application, 2021, 44(5): 187-192.

[6] MEHTA S, KARWA R, CHAVAN R. Keyphrase Extraction using Graph-based Statistical Approach with NLP Patterns [J]. Sādhanā, 2024(49).

[7] SUN Y, QIU H, ZHENG Y, et al. SIFRank: A New Baseline for Unsupervised Keyphrase Extraction Based on Pre-Trained Language Model [J]. IEEE Access, 2020(8): 10896-10906.

[8] GROOTENDORST M, WARMERDAM V D. MaartenGr/KeyBERT: V0. 5[EB/OL]. [2024-01-31]. https://github.com/MaartenGr/KeyBERT.

[9] SORNLERTLAMVANICH V, YUENYONG S. Thai Named

Entity Recognition using BiLSTM-CNN-CRF Enhanced by TCC [J]. IEEE Access，2022(10)：53043-53052.

[10] IZUTSU J，KOMIYA K，SHINNOU H. Word Segmentation of Hiragana Sentences Using Hiragana BERT[C]// Trends in Artificial Intelligence(PRICAI 2023). 2023：323-335.

[11] YOSHINAGA N. Back to Patterns：Efficient Japanese Morphological Analysis with Feature-Sequence Trie[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg，PA：ACL，2023：13-23.

[12] 京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所. MeCab：Yet Another Part-of-Speech and Morphological Analyzer [EB/OL]. [2024-07-20]. http://taku910. github. io/mecab/.

[13] Matthew Honnibal，Ines Montani. spaCy [EB/OL]. [2024-07-20]. https://github. com/explosion/spaCy.

[14] KAWANAMI S，HIDEMA K，OKADA K. Proposal of a Method Extracting Strategic Phrases from Japanese Enterprise Disclosure Documents[C]// Proceedings of the 9th International Congress on Advanced Applied Informatics. Piscataway，NJ：IEEE，2020：506-511.

[15] KIRIHARA T，MATSUMOTO K，YOSHIDA M，et al. Keyword extraction and classification from TV program viewers' tweets[C]// Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence. Tokyo：JSAI，2020：360-369.

[16] TANAKAA R，NAKAYAMAB S. Extraction of Chemical Substance Names from Patent Publication[J]. Journal of Computer Chemistry，2022(21)：1-9.

[17] HADIFAR A，STERCKX L，DEMEESTER T，et al. A Self-Training Approach for Short Text Clustering [C]// Proceedings of the 4th Workshop on Representation Learning for NLP. Stroudsburg，PA：ACL，2019：194-199.

[18] GUAN R，ZHANG H，LIANG Y，et al. Deep Feature-Based Text Clustering and its Explanation [J]. IEEE Transaction on Knowledge and Data Engineering，2022，34(8)：3669-3680.

[19] PUGACHEV L，BURTSEV M. Short Text Clustering with Transformers [EB/OL]. (2021-01-31) [2024-01-31]. https://arxiv. org/pdf/2102. 00541. pdf.

[20] ZHANG D，NAN F，XIAOKAI W，et al. Supporting Clustering with Contrastive Learning [C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Stroudsburg，PA：ACL，2021：5419-5430.

[21] AO Z，WENNING H，GANG C，et al. DEC-transformer：deep embedded clustering with transformer on Chinese long text [J]. Pattern Analysis and Applications，2023(26)：1349-1362.

[22] LIU E，IZUMI K，TSUBOUCHI K，et al. Cross-lingual news article comparison using Bi-graph Clustering and Siamese-LSTM [C]// Proceedings of the 31st Annual Conference of the Japanese Society for Artificial Intelligence. Tokyo：JSAI，2017：52-57.

[23] THANG D T，IWAI C，ONISHI K. A keyword clustering system based on search motivation for search marketing with BERT and HDBSCAN [C]// Proceedings of the 86th National Convention of IPSJ. Tokyo：IPSJ，2022：85-86.

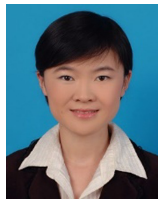
[24] SUZUKI M，SEKIZAKI N，KURODA S，et al. An Analysis on the Customer Logistic Satisfaction based on Word Clustering [J]. Innovation and Supply Chain Management，2023，17(1)：11-16.

[25] 国立国語研究所言語資源開発センター.「UniDic」国語研短単位自動解析用辞書[EB/OL]. [2024-7-20]. https://clrd. ninjal. ac. jp/unidic/.

[26] CHEN J Y. Research on Chinese Text Similarity Detection Technology Based on Word Weight Analysis[D]. Zhengzhou：Zhengzhou University，2021.

[27] NHN Japan 株式会社. livedoor ニュースコーパス[EB/OL]. [2024-07-20]. https://www. rondhuit. com/download. html # ldcc.

[28] The National Institute of Information and Communications Technology(NICT). Japanese Wiki Corpus Generated from the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles [EB/OL]. [2024-07-20]. https://www. japanesewiki. com/.



YU Juan，born in 1981，professor，Ph.D supervisor. Her main research interests include data science and intelligent information system.

LI Weiting，born in 2000，postgraduate. Her main research interests include data mining and business intelligence.