arXiv:2302.01005v1 [cs.CL] 2 Feb 2023

食物概念与食谱的预定义领域特定嵌入：异构食谱数据集案例研究

Gordana Ispirova[1,2]约瑟夫·斯特凡研究所计算机系统系，斯洛文尼亚卢布尔雅那[2]约瑟夫·斯特凡国际研究生院，斯洛文尼亚卢布尔雅那 gordana.ispirova@ijs.si

Tome Eftimov[1,3]约瑟夫·斯特凡研究所计算机系统系，斯洛文尼亚卢布尔雅那 tome.eftimov@ijs.si

Barbara Korou˘si´c Seljak[1,4]约瑟夫·斯特凡研究所计算机系统系，斯洛文尼亚卢布尔雅那 barbara.korousic@ijs.si

摘要——尽管如今食谱数据很容易获取，但要找到完整的食谱数据集却十分困难——其中需要包含配料表、每种食材及每道食谱的营养素含量、过敏原等。食谱数据集通常收集自用户发布和分享食谱的社交媒体网站。食谱通常几乎没有结构，而且计量单位既有标准化的，也有非标准化的。我们收集了六个不同的公开食谱数据集，它们采用不同格式，其中一些还包含不同语言的数据。为了将这些数据集转换为所需格式，以便应用机器学习（ML）流程进行营养素预测 **[1], [2],** ，我们使用基于词典的命名实体识别（NER）、基于规则的命名实体识别，以及外部领域专用资源进行转换，从而实现数据规范化。我们根据配料表创建领域特定嵌入，并让所有食谱使用相同的嵌入空间——由此生成一个食材数据集。规范化处理的结果是两个语料库：一个包含预定义的食材嵌入，另一个包含预定义的食谱嵌入。我们在全部六个食谱数据集上评估了机器学习流程。该应用案例的结果还证实，采用领域启发式方法合并的嵌入比基线效果更好。

索引词——食谱嵌入、配料嵌入、预定义语料库、机器学习流程、领域知识、预测建模

I. 引言

在这个社交媒体时代，包含食谱数据的互联网内容随处可见。然而，可用于研究，或可在需要结构化数据的应用场景中重复使用的、包含食谱数据的结构良好数据集却十分罕见，几乎不存在。食谱数据集通常是从向用户提供食谱的以食品为主题的在线社交网络服务、网站和移动应用中收集并提取的；这些食谱由其他用户撰写。此类服务不胜枚举，例如 Allrecipes [3], Yummly [4], Epicurious [5]。这些食谱通常几乎没有结构，使用的既有标准计量单位，也有非标准计量单位。

Recipe1M 是唯一公开可用的食谱数据集，包含结构化数据——分隔开的配料表、数量和度量单位，以及每份食谱和每种食材的营养素含量。在 [1] 中，我们介绍了一种机器学习流程（称为 P-NUT ），用于根据描述食品项目的文本所学习到的向量表示，预测该食品项目的营养素含量。基于这一机器学习流程，我们在 [2] 提出了一种用于合并文本嵌入的领域启发式方法。评估结果表明，在用于预测营养素含量的建模中，采用领域启发式方法合并的嵌入作为特征，优于采用传统合并技术合并的嵌入。Recipe1M 是唯一具备应用所提机器学习流程所需数据的公开可用食谱数据集。本研究旨在利用异构多语言数据集学习并生成食谱数据的预定义嵌入。

文本嵌入已广受关注，并成为许多自然语言处理研究中不可或缺的标准组成部分。其主要用途是“迁移学习”：先利用非常庞大的原始未标注数据集（来自不同来源，例如网页抓取）学习嵌入，然后将这些预训练的，即预定义的嵌入用作模型在监督任务中的输入（即经典的机器学习流程）。

在 [6] 中，作者展示了将迁移学习用于文本数据的能力。得益于预训练的计算机视觉模型 [9]—[11]，（这些模型均基于 ImageNet 数据库 [12]进行预训练），迁移学习已被证明在图像任务 [7], [8] 中表现出色。
在领域方面，医学领域已有类似研究，例如 cui2vec [13] ——一组为 108,477 个医学概念预先定义的嵌入。在[14] 中，作者介绍了 Med2Vec——一种简单、稳健的算法，它利用包含超过一百万次就诊的大型电子健康记录数据集，学习代码和就诊表示。它使我们能够解读这些经临床专家验证为有效的表示。

然而，在食品与营养领域，缺乏这类研究。 [15],介绍了现有以文本内容为重点来学习食谱表示的研究

[16]。这些研究利用了食谱相关的烹饪说明和食材。另有一些研究考虑了图像 [17]–[20] ，通常专注于食谱图像检索任务，并尝试将图像和文本对齐到同一共享空间中，从而导致两种模态的信息都遭到损失。在 [21], 中，作者专注于利用从图像、文本和关系中提取的多模态信息来学习食谱表示，重点研究基于食谱类别的菜系分类和区域预测；这意味着其中包含的文本数据仅通过食谱类别体现。

在这一背景下，预测营养素含量的任务仍未得到探索。此外，在我们的工作 [2] 之前，从未有人提出过基于食材用量的食谱嵌入。本研究从六个异构食谱数据集收集食谱数据：印度食谱、Recipe1M、Epicurious、沙拉食谱、Yumlly28K、Recipe box。首先，通过命名实体识别（NER）和数据映射流程，将这些数据集统一为相同格式。我们使用基于词典的命名实体识别提取计量单位，并使用基于规则的命名实体识别提取食材用量。数据映射过程融合了多种方法，用于将食材映射到美国农业部（USDA）[22]的食品成分数据库（FCDB）FoodData Central USDA，以获取营养素含量。本研究的最终成果包括两个预定义嵌入语料库，分别用于食材向量表示和食谱向量表示；此外，还包括四个由领域专家参与构建并审核通过的领域词典：一个用于计量单位的词典、一个用于转换计量单位的词典、一个用于品牌食品名称的词典，以及一个专用于食谱数据的冗余词词典。针对特定任务训练嵌入是一个非常耗时的过程，因此，预定义嵌入语料库既可用于研究，也可迁移至其他任务以供应用。

本文其余部分的结构如下：第二节首先介绍相关工作，并详细说明实验数据和方法。第三节展示实验结果、评估和进一步讨论。最后，第四节总结该方法的重要性，并指出未来工作的方向。

II. 方法

本节首先回顾理解该方法所需的最新研究成果、所采用的算法以及该领域近期的工作；随后介绍实验所用的数据，最后详细说明该方法。

A. 相关工作

1) FoodEx2分类系统：FoodEx2 [23] 是一种全面的食物分类与描述方法，旨在满足多个食品安全领域数据集合对食物描述的需求。FoodEx2是一种用于对食物进行分类和表征的标准化方法，是标准样品描述（SSD2）[24] （用于描述食品和饲料样品及分析结果的数据模型）的补充。它由对众多不同食品项目的描述构成，这些项目按父子层级结构归入食物组和更大的食物类别。

2) StandFood：StandFood [25] 是一种依据FoodEx2分类系统对食物进行标准化的方法，由三个部分组成。第一部分采用机器学习方法，将食物分为四类FoodEx2类别：单一食物的原始食物（r）和衍生食物（d），以及复合食品（c）的简单食物（s）和汇总食物（a）。第二部分结合自然语言处理方法和概率论来描述食物。为了提高分类结果，StandFood方法的第三个组成部分通过制定后处理标准，整合第一和第二部分的结果。StandFood的评估结果表明，该系统取得了良好的效果，可用于依据 FoodEx2 对食品项目进行分类和描述。StandFood还可以帮助用户在食品成分数据库和食品消费数据中查找缺失的 FoodEx2代码，从而进行比较和合并。

3) 美国农业部食品成分数据库：FoodData Central是美国农业部（USDA）的食品成分数据库（FCDB），是一个以研究为导向的综合数据系统。除了提供有关营养素和其他食物成分的更多数据外，它还链接到农业、食物、膳食补充剂及其他主题的信息来源。近年来，由于食品供应的变化速度不断加快，且食物数据的用途日益多样化，人们对公开且易于获取的食物和食品中营养素及其他成分信息的需求显著增加。为了以严谨、科学的方式分析、收集和呈现膳食概况数据，需要一种新的方法。Food Data Central 正是这一创新策略的体现。美国农业部食品数据中心 FoodData Central USDA FDCB 中有四个数据集与我们的研究相关：

a. 基础食品 – 包含对单个商品及其衍生的最低限度加工食品样本进行营养素评估后确定的数值，并提供有关其变异性的信息以及重要的底层元数据。b. 膳食研究营养素数据库（FNDDS）– 关于 What We Eat in America 中报告的食物和饮料的营养素及份量重量的数据；该调查属于全国健康与营养调查（NHANES） [26] 的膳食摄入部分。c. 品牌食品 – 来自公私合作项目收集的国内和国际品牌食品标签的数据。d. 标准参考数据库（SR）传统版 – 关于食物成分的历史数据，包括通过分析、计算和已发表文献得出的营养素数据。

4) 领域词典：为了从食谱数据集的文本中提取所需信息，需要使用领域词典，具体来说，是包含计量单位的词典（涵盖国际单位制（SI）和家用计量单位）。遗憾的是，当时没有适用于我们研究的现成词典，因此我们在一位领域专家——营养师的协助下，着手汇集所需资源。

a. 计量单位词典：该词典在一位领域专家——营养师的协助下编制，收录了食谱数据中不同食品项目可能使用的计量单位：

– 国际单位制（SI）中的数量计量单位——这些计量单位是官方单位，在 [27], [28], 中定义，并由国际计量大会（CGPM） [29]制定和维护。示例：克、毫克等。

– 家庭计量系统中的数量计量单位——该系统并非官方计量单位体系，但在日常生活中最为常用。示例：汤匙、茶匙、杯等。

– 药衡制中的数量计量单位——药衡制最初是用于配药和开处方的重量与计量体系，是一种历史上的质量和体积单位体系，曾由医生和药剂师用于开具处方，有时也被科学家使用 [30]–[32]。示例：品脱、打兰、斯克鲁普尔等。

计量单位词典包含 144 个实例。

b. 冗余词词典：该词典列出了一系列在此情境下可视为冗余的词语和短语。这些内容包括：对相关配料的营养成分没有任何实质影响或价值的词语，以及食谱撰写者添加的额外说明，用于进一步解释配料本身或烹饪过程。该词典共包含 415 个词语和短语。冗余词/短语示例：“1 磅大红葱头，必要时将球茎分开，并将每个球茎纵向切成两半”，“12 盎司品质最佳的三文鱼，切成薄如纸片的薄片，分开排列在容器中，然后放入冰箱冷藏半小时”。

c. 品牌食品词典：该词典收录了常见的食品品牌名称，这些名称通常可以替代其所生产的实际食品项目的名称。该词典从 USDA FCDB 的品牌食品数据集中提取。此类食品项目的示例包括：“M&M's”、“可乐”、“甜菊糖”、“奥利奥”、“雀巢咖啡”、“胡椒博士”、“七喜”等。

d. 换算词典。

该词典根据上述计量体系、多个在线换算计算器以及专家的意见和建议编制而成。其目的是将所有数量换算为同一种计量单位——克。该词典包含计量单位词典中所有实例对应的换算结果。

5) P-NUT 利用文本描述预测营养素含量：P-NUT [1] 是一种新颖的机器学习流程，用于学习预测模型，并将外部语义资源中编码的领域知识纳入其中。该机器学习流程（P-NUT）由三部分组成：

a. 表征学习：Mikolov 等人在 2013 [33] 和 Pennington 等人在 2014 [34], 中引入的词嵌入，在过去几年里已成为自然语言处理（NLP）任务中不可或缺的工具，也使各种以向量表示为输入的机器学习模型得以受益于这些高质量的文本输入表示。这类表示保留了更多词语的语义和句法信息，因此成为 NLP 领域的先进方法。 b. 无监督机器学习：不同类型食物的营养素含量存在显著差异。在一个大型数据集中，包含来自各类食物组的生鲜／简单食物和复合／食谱食物，每 100 g 食物中营养素的含量可能在 0–100 g 之间。 c. 监督式机器学习部分：P-NUT 方法的最后一部分是监督式机器学习，针对我们希望预测的各种营养素分别训练预测模型。营养素含量属于连续数据；因此，模型采用单目标回归算法进行训练，输入为短文本描述的已学习嵌入，并根据所选／可用的特定领域标准进行聚类。为此选择合适的机器学习算法颇具挑战；普遍采用的方法是选取若干机器学习算法，设定超参数范围，进行超参数调优，并利用交叉折叠验证评估估计器的性能（每次迭代使用相同的数据），最后对算法进行基准比较并选出最佳算法。回归算法最常用的基线是集中趋势度量，即数据集中训练部分的均值和中位数，用于所有预测。

6）合并文本嵌入的领域启发式方法：基于我们在[1] 和 [35] 中提出的机器学习流程这一概念，我们构建了一个表征学习流程（见 [2]），旨在探究：与使用食谱描述的向量表示相比，改用配料表的嵌入时，预测结果会如何变化。一种食物的营养素含量取决于其

食材；因此，食材文本包含更多相关信息。我们定义了一种用于合并食材嵌入的领域启发式方法，将每种食材的数量结合起来，以便将其作为机器学习模型进行营养素预测时所使用的特征。实验结果表明，使用该领域启发式方法可以改善预测结果。用于蛋白质预测的预测模型表现出很高的有效性，准确率最高可达97.98%。采用领域启发式方法组合多词嵌入，比使用传统的合并启发式方法能够取得更好的结果，在某些情况下准确率最高提升60%。

表 I 食谱数据集中结构良好且未拆分的食材列表示例。

示例	2 杯全麦面粉； 15 克泡打粉； 300 毫升 3.5%牛奶； 2 汤匙可可粉。
数量列表	[2, 15, 300, 2]
单位列表	[杯、克、毫升、汤匙]
食材名称列表	[全麦面粉， 泡打粉， 3.5%牛奶， 可可粉]

B. 数据

本研究考察了所有公开可用的食谱数据集，并从中选出除 Recipe1M 外另外六个有望转换为所需格式的数据集。所选数据集（包括 Recipe1M）如下：

a. Recipe1M——包含 51,500 道食谱，每道食谱包含以下数据：食谱标题（对食谱的简短文本描述）、结构化食材列表、食谱说明、食材的营养素含量（每种食材每 100 克中脂肪、蛋白质、饱和脂肪、钠和糖类的克数）、每种食材的数量、每种食材的计量单位（家庭计量系统）、每种食材的克重、营养素含量（每 100 克食谱中脂肪、蛋白质、盐、饱和脂肪和糖类的克数）、每 100 克食谱对应的 FSA交通灯标签。b. 印度食谱——包含 6,871 道食谱，每道食谱包含以下数据：食谱网址、以印地语连续书写的原始文本，其中包括食材、数量和计量单位，以及食谱说明。c. Epicurious——包含 20,103 道食谱，每道食谱包含以下详细信息：食谱标题、食谱网址、以连续文本形式呈现的原始食材、数量和计量单位、热量含量，以及蛋白质、脂肪和钠的营养价值。d. 沙拉食谱——包含 82,243 道食谱，每道食谱包含以下详细信息：食谱标题、食谱说明、以连续文本形式呈现的原始食材、数量和计量单位。e. Yummly28K——包含 27,639 道食谱，每道食谱包含以下数据：食谱标题、以连续文本形式呈现的原始食材、详细的营养素信息——94 种营养素的营养素含量。f. RecipeBox——包含 39,802 道食谱，每道食谱包含以下数据：食谱标题、以连续文本形式呈现的原始食材、数量和计量单位，以及食谱说明。

C. 方法

1）数据规范化：收集数据集后，需要将其转换为所需格式，以便应用机器学习流程。该过程分为两个部分：提取所需信息——使用基于词典的命名实体识别提取计量单位，并使用基于规则的命名实体识别提取配料的数量；以及数据映射——将配料（食品项目）映射到美国农业部食品成分数据库。

a. 从非结构化食谱数据中提取信息——在采取任何步骤将数据集转换为所需格式之前，务必注意，其中两个数据集包含非英语数据——印度食谱和 Yummly28K。这意味着我们处理的是多语言数据。为了翻译这些数据，我们使用了 Python 中的谷歌翻译 API [36] 。由于处理的是单句文本，几乎没有上下文，因此通过这种方式获得的结果是可靠的。

现在，让我们定义应用机器学习流程所需的理想格式。理想情况下，我们需要一个数据集 D ，其中包含 n 个实例/食谱，以及 m个用于表示这些实例的特征。在这些m 特征中，我们尤其关注以下内容：

– 食谱名称；– 整份食谱的营养信息（按 100 克计）；– 食材名称列表；– 每种食材以克为单位的数量列表（理想情况下按 100 克计）；

遗憾的是，没有任何数据集采用这种格式，只有 Recipe1M 将食材名称和数量分别列出。其他五个数据集都将食材、数量和单位以连续的原始文本形式呈现。因此，第一步是将每个数据集中每份食谱的食材名称、数量和单位分别提取到三个独立的列表中。如果文本遵循某种结构，这个过程只需简单地拆分文本（即拆分字符串），即可通过 Python 中的简单函数 [37] 和使用正则表达式搜索模式[38]，[39]，来完成。表 I 中给出了一个示例，展示了这一过程的实现方式。

检查数据集，尤其是包含配料表、数量和单位的原始连续文本时，很明显，这种技术

在这种情况下并不足够，因为文本不遵循任何模式或规则。我们感兴趣的所有食谱数据集中的文本都没有任何结构。有几点需要考虑：

– 计量单位：
  * 类型可能有所不同：
    · 国际单位制（SI）中的单位，即公制单位（例如克、毫升）；
    · 家用计量单位（例如杯、汤匙）；
    · 近似计量单位（例如少许、一撮、少量、一把、拇指大小、极少量）；
    · 特定食物或一组食物所用的单位（例如“一瓣大蒜”、“一颗生菜”、“一条肉桂”或“一条黄油”）。
  * 可以有不同的写法：
    · 使用完整拼写（例如：克、汤匙）；
    · 使用不同类型的缩写（例如：汤匙可缩写为 tb、tbs、tbsp；磅可缩写为 lb、lbs、lbss）；
  * 也可以不存在（例如：“3 个鸡蛋”、“2 个洋葱”）
– 数量：
  * 可以有许多不同的写法：
    · 用数字表示（例如“250 毫升橙汁”）；
    · 用文字表示（例如“两个苹果”）；
    · 数字和文字混合使用（例如“4 汤匙半糖”、“两小包 0.16 盎司方便面”）；
    · 使用不定冠词（例如“一杯牛奶”）；
  * 也可以不存在（例如“盐和胡椒”）。
  * 也可以由“用户自定义”（例如“肉桂，适量”）。
– 表示配料的文本：
  * 可以是简单的食物（例如“2 磅樱桃”）；
  * 可以是品牌食品（例如“1 小包（40 克）M&M’s 巧克力糖果）；
  * 可以包含有关食物本身的补充说明（例如“1 磅大红葱头（8 个），如有需要，将球茎分开，并把每个球茎纵向切成两半”“1⁄2 磅去皮鸡胸肉（约 1 块鸡胸肉，切成 1⁄2 英寸见方的小块））；
  * 可以是非食物，但在食谱中需要用到，因此会写在食材列表里（例如“1 片烘焙纸”）。

最初的尝试是将所有这些情况归纳起来，并制定字符串拆分规则，包括提取数量和计量单位的规则。

– 提取计量单位 – 为了从文本中提取计量单位，我们构建了一个在 II-A4小节中定义的词典——因此，从该词典中识别到的文本实体都被定义并提取为计量单位。

– 提取数量 – 为了从文本中提取数量，我们定义了若干规则和模式：

  *如果存在以下格式的数字出现次数 $D_{integer}\frac{D_{integer}}{D_{integer}}$. 其中 D 可以是一个或多个 – 位整数。这些数会合并为一个数字。
  * If 那里 are 数字 出现次数 in the 格式： $D_{integer}$ $D_{integer}$, $D_{integer}$ D_{float}, D_{float} $D_{integer}$, $\frac{D_{integer}}{D_{integer}}$ $\frac{D_{integer}}{D_{integer}}$, $\frac{D_{integer}}{D_{integer}}$ $D_{integer}$, D_{float} D_{float}。 其中 $D_{integer}$ 可以是一个或多个数字的整数数字，而 D_{float} 是两位或多位的浮点数数字。如果出现这种情况，将该出现项中的数字相乘。

– 提取食品项目 – 为了从提取计量单位和数量后剩余的文本中提取食品项目，我们使用冗余词词典，并删除词典中所有词语的出现项。剩下的内容就是食品项目。

有一些特殊情况：某些词在某些情况下可视为计量单位，而在其他情况下则可省略。当以下任一词语：“罐”、“罐头”、“小包”、“包装”、“盒”、“瓶”和“容器”出现在表示配料的文本中时，可能出现两种情况：

– 如果其后紧跟定义的计量单位词典中的计量单位，则省略该词（例如：“1 (500 克) 容器希腊酸奶”）。– 如果在未提取出定义词典中的任何计量单位的文本中发现该词，则将其视为计量单位（例如：“2 小包甜菊糖”）。

某些食品项目有特定的计量单位，这些单位只对应这些食品项目。只有当这些词与其对应的食品项目成对出现时，才会将其提取为计量单位。这些词对包括：

– “瓣”和“大蒜”；– “条”和“黄油”/“人造黄油”/“肉桂”/“胡萝卜”/“芹菜”；– “枝”和“迷迭香”/“百里香”/“薄荷”/“欧芹”；– “节”和“香肠”；– “茎”和“芹菜”/“青葱”/“小葱”/“西兰花”/“羽衣甘蓝”/“花椰菜”；– “片”和“明胶”；– “块”和“高汤”/“黄油”/“人造黄油”；– “颗”和“卷心菜”/“生菜”/“花椰菜”；

当然，这些成分对于不同食品项目的权重各不相同，换算词典中已考虑到这一点。

因此，我们可以看到，提取计量单位的过程是一个结合命名实体识别的混合过程

并辅以词典和额外规则。应用这些规则后，应能分别得到计量单位（一个或多个）、数量（一个或多个）和食品项目的列表。然而，在定义这些步骤后，我们发现提取所需信息的过程还存在更多困难。具体来说，配料表中指定给一种配料的文本可能包含多种配料（用连词“加”“和”或“与”连接）、同一种配料的多个选项（用连词“或”连接）、同一种配料的多个数量（用介词“至”连接），或上述两种及以上情况的组合。针对这些情况，适用的解决方案如下：

– 包含连词”and”或”with”：

∗如果满足以下条件：($count$(”and”) $\geq$ 1 $\vee$ $count$(”$with$”) $\geq$ 1) $\wedge$ $count$(”$plus$”) == 0 连词两侧各提取出两个计量单位，则将连词”and”替换为”plus”，并在后续处理中按”plus”处理（例如：”2 4 盎司包装的三文鱼和 5 汤匙柠檬汁”）。

∗ 其他包含连词”and”的情况均予以忽略（例如：”1 磅大红葱头（8 个），必要时将球茎分开，并将每个球茎纵向切成两半”）。– 包含连词”plus”：

∗ 如果 $count$(”$plus$”) $\geq$ 1 – 按”plus”拆分，并根据文本中”plus”的数量将文本拆分为相应数量的配料（例如：”1/4 杯切碎的甜酸小黄瓜，加 2 汤匙罐中的腌渍汁，再加 12 根整根小黄瓜”）。– 包含连词”or”：

∗ 如果 $count$(“or”) $>$ 1 – 按“或”拆分，并将文本分成两部分，但仅保留第一部分；

∗ 如果 $count$(“or”) $>$ 1 – 检查所提取的计量单位的位置，根据“或”的数量拆分成相应数量的部分，并：

· 如果提取出一个计量单位，则保留包含该计量单位的部分。

· 如果提取出多个计量单位，且所有或没有计量单位被分配了数量，同时配料中没有“加”，则保留包含第一个计量单位的部分。

· 如果提取出多个计量单位，其中一些未被分配数量，且配料中没有“加”，则保留包含已分配数量的计量单位的部分。– 包含介词“到”：

∗ 如果“to”出现在特定模式中，且该模式包含整数和/或浮点数和/或以“/”分隔的数字（例如：“2 14 1/2- to 15-ounce cans diced tomatoes in juice”），则按“to”拆分该模式后，保留前半部分并将其附加到文本其余部分。

尽管如此，我们仍发现还有其他情况：指定为一种食材的文本中出现了两种食材。更具体地说，这些文本包含以下任意短语：“mixed with”、“mixed in”、“beaten with”、“dissolved in”、“sauteed in”、“diluted”、“combined”，或动词“add”。解决方法是将这些短语视为“plus”并替换为该词。例如：“5 to 6 凤尾鱼 add 1 茶匙红辣椒碎”、“1 large 鸡蛋 beaten with 一撮盐”、“1 茶匙玉米淀粉 dissolved in 2 汤匙牛奶”。另一个反复出现的问题是同一种食材对应多个数量，也就是说，文本中包含了多余的信息。这种情况会在多种情形下出现，我们针对每种情形都采用了相应的解决方案：

– 如果我们提取出两个表示数量的不同模式，并且第二个模式后有一个计量单位，且两个表示数量的模式之间没有文本 – 保留第二个模式作为数量，并从文本中删除第一个模式。– 如果我们提取出两个表示数量的不同模式，并且第一个模式后有一个计量单位，紧接着就是第二个模式 – 保留第一个模式作为数量，并删除第二个模式。– 如果我们提取出两个表示数量的不同模式和两个不同的计量单位，并且两个模式之间没有文本，只有与第一个模式相连的计量单位 – 保留第一个模式作为配料的数量。– 如果我们提取出两个表示数量的不同模式，且没有提取出计量单位，并且两个模式之间没有文本 – 保留第一个模式作为配料的数量。

在应用任何规则之前，会先对文本进行预处理。预处理遵循以下几条规则：

– 对文本进行分词和词形还原。– 删除括号以及冗余词词典中的冗余词。– 识别品牌食品词典中的词，并且不对这些词进行进一步预处理。– 删除特殊字符，但以下情况除外：

∗ ”%” – 如果前面有数字。

∗ ”&” – 将其替换为”and”。

∗ ”+” – 将其替换为”plus”。

∗ ”/” – 如果它位于两个数字之间。

∗ ”-”在数字之间时，会替换为”to”。–表示数字的词语会替换为其代表的数字（特殊情况包括将”half a”、”half of”、”half from”替换为”1/2”）。– 如果文本以”a/an”开头，后面紧跟一个计量单位，则将”a/an”替换为”1”。–如果文本以”a/an”开头，且不存在计量单位，也不存在某种模式

表示文本中存在数量的模式，则将”a/an”替换为”1”。–如果提取出的数量和计量单位之后出现”.”或”,”，且文本中没有”and/or”，则删除句点或逗号之后的所有内容。

将这些处理应用于数据集中的所有实例后，我们便得到六个新数据集，其格式适用于生成嵌入——即 [2] 中所述方法的表示学习部分。

b. 数据映射至美国农业部食品成分数据库——为了应用监督式机器学习部分，我们首先需要确定要预测哪些营养素含量。从 II-B 小节中给出的数据集说明可以看出，只有三个数据集包含营养素含量，而且包含的营养素完全相同。为了获取所有数据集的营养素含量，我们选择将数据映射至食品成分数据库，具体选用美国农业部食品成分数据库 [22], 。这一选择不仅因为这些数据集大多来自美国的网站，也因为该数据库是规模最大的综合数据系统，能够提供更详尽的营养成分数据，并链接至相关农业和实验研究。该数据库有两个用途：

1) 获取数据集中缺失的以克为单位的数量——这些数量之所以缺失，是因为没有提供计量单位，只有食品项目和/或一个数字（例如：“2 个鸡蛋”、“3 个苹果”、“1 个标准大小的鸡胸肉”等）。2) 获取每种配料的营养素含量，以便计算食谱的营养素含量。

在数据映射过程之前，我们构建了一个新的数据集，其中包含从全部六个食谱数据集中提取的所有唯一配料。该数据集总共包含 71,641 个实例。通过将食材映射到 USDA 数据库中的食品项目，我们实现了上述两个目标。我们在以往的研究中曾处理过此类映射 [35], [40], 当时采用了涉及词性标注和概率论的词汇相似度方法。由于该方法基于名词交集进行匹配，因此存在局限性，而词性标注方法可能无法捕捉到这些名词。此外，实例可能不包含任何名词，或者词性标注方法可能无法正确标注词元（单词）（例如，“orange”可能被标注为形容词，而不是名词）。在本研究中，为尽可能多的食材获得匹配结果至关重要，因此我们不能仅依赖这种方法。我们确实将该方法纳入了数据映射中——我们纳入名词、动词、形容词和数字，并将相似度重新表述为

如方程（1）、（2）和（3）所示。

$$p_n = \frac{Nouns_i \cap Nouns_j}{Nouns_i \cup Nouns_j} \tag{1}$$

$$\begin{aligned} p_{v+a+num} = \big(((Verbs_i \cup Adjectives_j \cup Numbers_i) \cap \\ (Verbs_j \cup Adjectives_j \cup Numbers_j)) + 1\big) \\ / \ \big(((Verbs_i \cup Adjectives_j \cup Numbers_i) \cup \\ (Verbs_j \cup Adjectives_j \cup Numbers_j)) + 2\big) \end{aligned} \tag{2}$$

$$similarity_{index} = p_n \times p_{v+a+num} \tag{3}$$

我们希望匹配的配料中的名词、动词、形容词和数字集合为：$Nouns_i$、$Verbs_i$、$Adjective_i$、$Numbers_i$；而来自 USDA 的食品项目中的相应集合为：$Nouns_j$、$Verbs_j$、$Adjective_j$、$Numbers_j$。

为了执行从美国农业部食品成分数据库进行的映射，我们从 USDA 的四个数据集中提取了食品项目。初始合并数据集包含 298,318 个项目；移除重复项后，最终合并数据集包含 290,519 个食品项目。随后，该数据集经过一些预处理：文本被分词并进行词形还原，括号以及冗余词词典中的冗余词被移除。

这两个数据集——包含六个食谱数据集中的唯一配料，以及来自美国农业部食品成分数据库四个数据集的预处理食品项目——随后作为数据映射过程的输入。映射过程是由多种相似度度量组成的集成方案，这些度量按重要性排序。若满足以下相似度度量之一（按重要性从高到低排列），则认为两个食品项目——一个来自 USDA，另一个是数据集中的配料——相匹配：

– 词元集合相同。– 名词集合相同，且 $similarity_{index} > 0$。– 名词的交集非空，且 $similarity_{index} > 0$。– 两个食品项目的名词、动词、形容词和数字的并集的交集非空，且 $similarity_{index} > 0$。– 两个食品项目拼接后的词元的莱文斯坦距离 [41] 小于待匹配配料的拼接词元长度。

III. 结果与讨论

本节评估所提出的方法，包括实验设置、获得的结果，以及对实验结果的讨论。

A. 预定义配料嵌入语料库

数据规范化过程的结果是六个具有结构化格式、映射至 FCDB 的统一格式的食谱数据集。在从

非结构化食谱数据集中提取信息后，构建了一个包含 71,641 种配料的数据集。基于该数据集，使用四种嵌入算法和以下参数生成食材嵌入：

a. Word2Vec——架构：连续词袋模型和 Skip-gram模型，向量维度：50、100 和 200，滑动窗口：2、3、5 和 10，以及组合单个词嵌入的启发式方法：求和和平均值。b. GloVe——向量维度：50、100 和 200，滑动窗口：2、3、5 和 10，以及组合单个词嵌入的启发式方法：求和和平均值。c. Doc2Vec——架构：分布式记忆段落向量和分布式词袋段落向量，向量维度：50、100 和 200，滑动窗口：2、3、5 和 10，以及组合单个词嵌入的启发式方法：求和和平均值。d. BERT——神经网络最后四层的组合方式：求和和拼接。

结果是一个预定义的食材嵌入语料库。

B. 预定义食谱嵌入语料库

将含有配料的数据集映射到美国农业部食品成分数据库后，提取其中的营养素含量信息，并将其添加到每种配料旁边，即每 100 克配料所含的营养素含量。缺失的数量则从美国农业部食品成分数据库中食品项目的份量（以克为单位）提取。因此，计算食谱嵌入所需的数据都已齐备，可以使用领域启发式方法进行计算。

为了生成食谱嵌入，我们首先计算食谱每 100 克中各配料的数量，并据此按比例换算营养素含量，从而得到食谱每 100 克所含的营养素含量。我们选择以下五种营养素作为目标值：脂肪、蛋白质、糖、饱和脂肪和钠。这一选择显而易见，因为具有营养素含量数据的三个数据集都包含其中全部或部分营养素。

我们使用预定义的食材嵌入语料库，以及通过映射到美国农业部食品成分数据库计算出的营养素含量，结合 [2] 中介绍的领域启发式方法，为六个食谱数据集生成食谱嵌入。采用同一食材嵌入数据集来生成领域特定的食谱嵌入，使它们都处于同一空间，即同一特征空间。最终得到六个食谱嵌入数据集，并将它们合并为预定义食谱嵌入语料库，其中包含 219,765 个食谱嵌入。

C. 预测建模

获得嵌入后，下一步是进行预测建模：使用四种不同类型的嵌入算法（Word2Vec、GloVe、Doc2Vec 和 BERT），通过六种不同类型的回归（线性回归、岭回归、套索回归、弹性网络、决策树、随机森林和神经网络回归）进行单目标回归，以预测五种营养素（脂肪、蛋白质、糖、饱和脂肪和钠）。总共训练了 3,660 个模型：1,440 个 Word2Vec 模型、720 个 GloVe 模型、1,440 个 Doc2Vec 模型和 60 个 BERT 模型。预测模型通过以下步骤进行评估：

a. 超参数调优——使用 Python 中 Scikit-learn 库的 [42]）：对线性回归、岭回归、套索回归、弹性网络和决策树回归使用 GridSearchCV（遍历所有参数组合）；对随机森林和神经网络使用 RandomizedSearchCV（从参数空间中抽取指定数量的候选参数），因为它们是更复杂的算法，而 GridSearchCV 的穷举搜索需要耗费更多执行时间。b. 使用超参数调优得到的最佳参数进行训练。c. 通过 K 折交叉验证估计预测误差。d. 计算 [1],[2], [35] 中定义的领域特定准确率，以及基线——均值和中位数。

为了比较结果，我们还使用未通过领域启发式方法进行合并而获得的食谱嵌入训练了模型。我们在表 II、IV、V、VI、V 和 VII 中分别展示了六个数据集（Recipe1M、印度食谱、Epicurious、沙拉食谱、Yummly28K 和 Recipe box）采用和未采用领域启发式方法所获得的结果。

IV. 结论

通过本研究，我们协调化了六个不同的异构食谱数据集的元数据，并为研究界提供了两个预定义领域特定嵌入语料库（一个包含食材嵌入，另一个包含食谱嵌入），以及四种不同的领域特定词典。这些成果是在领域专家——一位营养师的帮助下创建的，包括：计量单位词典、计量单位到克的换算词典、品牌食品词典，以及针对食谱数据文本中冗余词语的词典。

如前所述，多个领域以及食品与营养领域已经存在预定义的领域特定嵌入，但针对包含外部领域资源且采用领域启发式方法进行合并的食谱数据，尚无此类嵌入。预定义的食材嵌入语料库目前是现有最大的食谱食品项目语料库，这些食品项目均作为配料出现在食谱中。该语料库作为独立的预定义嵌入语料库，可轻松用于并迁移至其他机器学习任务和研究中，而不仅限于本文介绍的任务——预测营养素含量。预定义的食谱嵌入语料库是最大的食谱嵌入语料库，包含针对 228,158 种不同食谱的领域特定嵌入，因此在食品与营养领域以及机器学习社区中都是一项强大的资源，因为针对特定机器学习任务训练新的嵌入既耗时又费力。这些嵌入与食材嵌入语料库一样，可以迁移至其他机器学习任务，并用于其他研究。

我们还评估了机器学习流程和领域启发式方法在合并多词嵌入方面的有效性。结果表明，机器学习流程不受数据集偏差影响，且能够处理异构数据。这些数据的多样性

表 II RECIPE1M 数据集的结果。

嵌入	目标	平均值准确率 (DH)	Mean 平均值准确率	中位数平均值准确率	平均值准确率（无 DH）
Word2Vec	Fat	67.86	7.86	7.48	21.88
	蛋白质	86.22	27.14	22.89	62.17
	饱和脂肪	85.35	7.27	8.78	49.69
	糖类	60.76	6.96	13.47	25.73
	钠	91.00	25.42	43.58	81.14
GloVe	Fat	62.96	7.84	7.53	21.46
	蛋白质	83.95	27.06	22.89	60.91
	饱和脂肪	81.69	7.27	8.78	49.54
	糖类	61.5	6.94	13.47	26.13
	钠	90.5	25.39	43.57	81.06
Doc2Vec	Fat	62.53	7.84	7.53	20.57
	蛋白质	82.6	27.06	22.89	59.76
	饱和脂肪	81.9	7.27	8.78	48.05
	糖类	55.62	6.94	13.47	22.47
	钠	89.6	25.39	43.57	80.93
BERT	Fat	77.64	7.84	7.53	26.29
	蛋白质	89.32	27.06	22.89	63.57
	饱和脂肪	90.95	7.27	8.78	55.08
	糖类	85.64	6.94	13.47	33.44
	钠	92.06	25.39	43.57	81.19

表 III
印度食谱数据集的结果。

嵌入	目标	平均值准确率 (DH)	Mean 平均值准确率	中位数平均值准确率	平均值准确率（无 DH）
Word2Vec	Fat	67.77	22.52	21.14	47.82
	蛋白质	81.4	43.07	29.82	70.29
	饱和脂肪	85.85	15.92	27.36	75.71
	糖类	59.88	8.69	21.53	36.67
	钠	92.88	95.63	47.01	84.88
GloVe	Fat	67.14	22.52	21.14	47.04
	蛋白质	81.07	43.07	29.82	69.99
	饱和脂肪	85.56	15.92	27.36	75.54
	糖类	59.44	8.69	21.53	36.09
	钠	92.88	95.63	47.01	84.87
Doc2Vec	Fat	66.08	22.52	21.14	46.60
	蛋白质	83.16	43.07	29.82	69.87
	饱和脂肪	85.02	15.92	27.36	75.24
	糖类	60.98	8.69	21.53	32.71
	钠	92.72	95.63	47.01	84.88
BERT	Fat	74.42	22.52	21.14	48.00
	蛋白质	86.61	43.07	29.82	71.00
	饱和脂肪	87.69	15.92	27.36	75.44
	糖类	75.86	8.69	21.53	35.91
	钠	92.78	95.63	47.01	84.87

也为将预测模型推广到许多不同场景奠定了基础，而且所需数据量极少。

致谢

本研究得到斯洛文尼亚研究局（核心研究资助项目编号 P2-0098）以及欧盟地平线2020研究与创新计划（FNS-Cloud，食品与营养安全）（资助协议编号 863059）的支持。

参考文献

[1] G. Ispirova、T. Eftimov 和 B. Korouˇsi´c Seljak，《P-NUT：根据简短文本描述预测营养素含量》，Mathematics，第 8 卷，第 10 期，第 1811 页，2020 年，出版商：多学科数字出版研究所。[2] ——，《用于营养素价值预测的多词嵌入领域启发式融合》，Mathematics，第 9 卷，第 16 期，第 1941 页，2021 年，出版商：多学科数字出版研究所。

表 IV EPICURIOUS食谱数据集的结果。

嵌入	目标	平均值准确率 (DH)	Mean 平均值准确率	中位数平均值准确率	平均值准确率（无 DH）
Word2Vec	Fat	54.04	19.22	12.89	33.06
	蛋白质	78.84	24.73	16.48	54.0
	饱和脂肪	86.05	18.92	16.03	75.96
	糖类	61.96	7.96	17.52	40.48
	钠	90.52	43.66	44.99	82.34
GloVe	Fat	52.77	19.1	12.89	33.25
	蛋白质	75.88	24.71	16.48	51.06
	饱和脂肪	86.03	18.9	16.04	76.04
	糖类	55.75	7.95	17.57	35.76
	钠	91.12	43.61	45.0	82.95
Doc2Vec	Fat	47.12	19.1	12.89	31.5
	蛋白质	66.66	24.71	16.48	48.36
	饱和脂肪	84.34	18.9	16.04	75.66
	糖类	45.0	7.95	17.57	30.5
	钠	90.56	43.61	45.0	82.84
BERT	Fat	59.32	19.1	12.89	31.71
	蛋白质	79.82	24.71	16.48	51.47
	饱和脂肪	86.93	18.9	16.04	76.09
	糖类	60.85	7.95	17.57	37.56
	钠	89.98	43.61	45.0	82.14

表 V 沙拉食谱数据集的结果。

嵌入	目标	平均值准确率 (DH)	Mean 平均值准确率	中位数平均值准确率	平均值准确率（无 DH）
Word2Vec	Fat	44.98	17.25	12.33	29.11
	蛋白质	69.79	21.91	16.0	49.84
	饱和脂肪	83.81	16.48	13.22	74.12
	糖类	50.38	8.75	17.55	34.99
	钠	88.06	22.39	44.85	76.01
GloVe	Fat	44.02	17.25	12.33	28.44
	蛋白质	66.96	21.91	16.00	46.39
	饱和脂肪	83.46	16.48	13.22	74.30
	糖类	45.67	8.75	17.55	32.49
	钠	89.95	22.39	44.85	82.09
Doc2Vec	Fat	40.56	17.25	12.33	26.21
	蛋白质	59.23	21.91	16.0	41.66
	饱和脂肪	82.44	16.48	13.22	73.5
	糖类	38.43	8.75	17.55	26.66
	钠	90.76	22.39	44.85	82.46
BERT	Fat	46.6	17.25	12.33	29.25
	蛋白质	70.66	21.91	16.0	48.9
	饱和脂肪	84.56	16.48	13.22	74.69
	糖类	51.86	8.75	17.55	33.2
	钠	89.92	22.39	44.85	80.54

表 VI Yummly28K 食谱数据集的结果。

嵌入	目标	平均值准确率 (DH)	Mean 平均值准确率	中位数平均值准确率	平均值准确率（无 DH）
Word2Vec	Fat	53.47	19.44	13.81	33.83
	蛋白质	76.51	27.06	15.33	51.85
	饱和脂肪	85.69	19.06	14.54	75.80
	糖类	64.17	9.62	28.34	42.43
	钠	89.64	26.18	45.83	77.73
GloVe	Fat	51.6	19.4	13.82	33.37
	蛋白质	73.69	27.09	15.33	49.38
	饱和脂肪	85.26	19.06	14.55	75.79
	糖类	58.54	9.64	28.35	40.03
	钠	88.25	27.03	45.84	79.46
Doc2Vec	Fat	44.79	19.4	13.82	31.71
	蛋白质	60.52	27.09	15.33	45.84
	饱和脂肪	83.54	19.06	14.55	75.15
	糖类	44.6	9.64	28.35	34.12
	钠	87.26	27.03	45.84	79.91
BERT	Fat	57.36	19.4	13.82	33.20
	蛋白质	77.26	27.09	15.33	48.26
	饱和脂肪	86.58	19.06	14.55	75.80
	糖类	65.31	9.64	28.35	38.82
	钠	82.92	27.03	45.84	75.98

表 VII Recipe Box 数据集的结果。

嵌入	目标	平均值 准确率 **(DH)**	**Mean** 平均值 准确率	中位数 平均值 准确率	平均值 准确率 （无 DH）
Word2Vec	Fat	59.27	18.61	11.08	35.11
	蛋白质	80.68	26.53	18.82	58.32
	饱和脂肪	83.79	16.00	10.04	72.30
	糖类	59.02	9.61	12.82	35.48
	钠	89.88	38.39	45.77	81.02
GloVe	Fat	56.51	18.61	11.08	34.50
	蛋白质	78.3	26.53	18.82	55.64
	饱和脂肪	83.09	16.00	10.04	72.04
	糖类	53.33	9.61	12.82	32.84
	钠	90.84	38.39	45.77	82.89
Doc2Vec	Fat	51.17	18.61	11.08	32.51
	蛋白质	70.94	26.53	18.82	51.94
	饱和脂肪	80.41	16.00	10.04	70.75
	糖类	42.91	9.61	12.82	27.24
	钠	90.13	38.39	45.77	82.22
BERT	Fat	67.43	18.61	11.08	26.29
	蛋白质	82.78	26.53	18.82	63.57
	饱和脂肪	86.64	16.00	10.04	55.08
	糖类	64.96	9.61	12.82	33.44
	钠	89.67	38.39	45.77	81.19

[3] Meredith Food Group。（2022）Allrecipes 网站。 [在线]。网址：https://www.allrecipes.com/ [4] （2022）Yummly 网站。 [在线]。网址：https://www.yummly.com/ [5] （2022）Epicurious 网站。[在线]。网址：https://www.epicurious.com/ [6] J. Howard 和 S. Ruder，《用于文本分类的通用语言模型微调》，arXiv 预印本 arXiv:1801.06146，2018 年。[7] V. Gulshan、L. Peng、M. Coram、M. C. Stumpe、D. Wu、A. Narayanaswamy、S. Venugopalan、K. Widner、T. Madams、J. Cuadros 等，《用于检测视网膜眼底照片中糖尿病视网膜病变的深度学习算法的开发与验证》，《美国医学会杂志》，第 316 卷，第 22 期，第 2402–2410 页，2016 年。[8] A. L. Beam 和 I. S. Kohane，《将人工智能应用于临床护理》，《美国医学会杂志》，第 316 卷，第 22 期，第 2368–2369 页，2016 年。[9] K. Simonyan 和 A. Zisserman，《用于大规模图像识别的超深卷积网络》，arXiv 预印本 arXiv:1409.1556，2014年。[10] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna，《重新思考用于计算机视觉的 Inception 架构》，载于 IEEE 计算机视觉与模式识别会议论文集，2016 年，第 2818–2826 页。[11] K. He、X. Zhang、S. Ren 和 J. Sun，《用于图像识别的深度残差学习》，载于 IEEE 计算机视觉与模式识别会议论文集，2016 年，第 770–778 页。[12] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei，《ImageNet：大型分层图像数据库》，载于 2009 年 IEEE 计算机视觉与模式识别会议。IEEE，2009 年，第 248–255 页。[13] A. L. Beam、B. Kompa、A. Schmaltz、I. Fried、G. Weber、N. Palmer、X. Shi、T. Cai 和 I. S. Kohane，《从海量多模态医疗数据中学习临床概念嵌入》，载于 2020 年太平洋生物计算研讨会。World Scientific，2019 年，第 295–306 页。[14] E. Choi、M. T. Bahadori、E. Searles、C. Coffey、M. Thompson、J. Bost、J. Tejedor-Sojo 和 J. Sun，《医疗概念的多层表示学习》，载于第 22 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，2016 年，第 1495–1504 页。[15] D. Li 和 M. J. Zaki，《Reciptor：一种用于食谱表示学习的有效预训练模型》，载于第 26 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，2020 年，第 1719–1727 页。[16] Y. Tian、C. Zhang、R. Metoyer 和 N. V. Chawla，《利用网络进行食谱表示学习》，载于第 30 届 ACM 信息与知识管理国际会议论文集，2021 年，第 1824–1833 页。[17] M. Carvalho、R. Cad`ene、D. Picard、L. Soulier、N. Thome 和 M. Cord，《烹饪场景中的跨模态检索：学习文本-图像语义嵌入》，载于第 41 届 ACM SIGIR 信息检索研究与发展国际会议，2018 年，第 35–44 页。

[18] A. Salvador、N. Hynes、Y. Aytar、J. Marin、F. Ofli、I. Weber 和 A. Torralba，“为烹饪食谱和食品图像学习跨模态嵌入”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第3020–3028页。[19] H. Wang、D. Sahoo、C. Liu、E.-p. Lim 和 S. C. Hoi，“利用对抗网络为烹饪食谱和食品图像学习跨模态嵌入”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2019年，第11 572–11 581页。[20] A. Salvador、E. Gundogdu、L. Bazzani 和 M. Donoser，“借助层次化Transformer和自监督学习改进跨模态食谱检索”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2021年，第15 475–15 484页。[21] Y. Tian、C. Zhang、Z. Guo、Y. Ma、R. Metoyer 和 N. V. Chawla，“Recipe2vec：利用图神经网络进行多模态食谱表示学习”，arXiv预印本 arXiv:2205.12396，2022年。[22] 美国农业部，USDA，食品成分数据库，2021年。 [在线]。网址：https://fdc.nal.usda.gov[23] 欧洲食品安全局，“FoodEx2食品分类与描述系统（修订版2）。” [在线]。网址：https://efsa.onlinelibrary.wiley.com/doi/epdf/10 2903 2015 804 。/2 0, 11, [24] sp.efsa. .EN- ——，“标准样品描述第 . 版。”《EFSA期刊》，第 . 卷，第10期，第3424页，2013年。[25] T. Eftimov、P. Koroˇsec 和 B. Korouˇsi´c Seljak，“StandFood：根据FoodEx2，利用半自动系统对食物进行分类和描述，以实现食物标准化”，《营养素》，第9卷，第6期，第542页，2017年，出版方：多学科数字出版研究所。[26] （2022年）全国健康与营养调查（NHANES）。[在线]。网址：https://www.cdc.gov/nchs/nhanes/index.htm [27] E. G´obel、I. Mills 和 A. Wallard，“国际单位制（SI）”，2006年。[28] D. B. Newell、E. Tiesinga 等，“国际单位制（SI）”，NIST特别出版物，第330卷，第1–138页，2019年。[29] J. Turner，“美国对国际单位制（公制计量系统）的解释”，《联邦公报》，第73卷，第96期，第28 432–3页。[30] J. F. Kr¨uger，《世界各国货币、质量和重量完整国际手册》.....G. basse，1830年。[31] W. Henry，《实验化学原理》。Robert Desilver，1831年，第2卷。[32] J. S. T. Gehler，《约翰·塞缪尔·特劳戈特·格勒的物理学词典》。EB Schurckert，1844年，第10卷，第3期。[33] T. Mikolov、I. Sutskever、K. Chen、G. S. Corrado 和 J. Dean，“词语和短语的分布式表示及其组合性”，载于《神经信息处理系统进展》，美国内华达州太浩湖，2013年12月，第3111–3119页。[34] J. Pennington、R. Socher 和 C. D. Manning，“GloVe：用于词表示的全局向量”，载于《2014年自然语言处理经验方法会议（EMNLP）论文集》，2014年，第1532–1543页。[35] G. Ispirova、T. Eftimov 和 B. Korouˇsi´c Seljak，“在食品与营养数据预测任务中探究知识领域偏差”，载于《2020年IEEE大数据国际会议（大数据）》。IEEE，2020年，第3563–3572页。[36] S. Han，“googletrans 3.0. 0”，PyPI库，2020年。[37] Python软件基金会，“Python 3.10.6文档”，2022年，（访问日期：2022年8月11日）。 [在线]。网址：https://docs.python.org/3/library/string.html [38] A. V. Aho，“字符串模式查找算法，《理论计算机科学手册》A卷”，A，J. van Leeuwen 编，Elsevier- SciencePublishersB，第1990卷，第257–297页，1990年。[39] R. Mitkov，《牛津计算语言学手册》。牛津大学出版社，2022年。[40] G. Ispirova、G. Cenikj、M. Ogrinc、E. Valenˇciˇc、R. Stojanov、P. Koroˇsec、E. Cavalli、B. Korouˇsi´c Seljak 和 T. Eftimov，“CafeteriaFCD语料库：针对不同食品语义资源进行标注的食品消费数据”，《食物》，第11卷，第17期，第2684页，2022年。[41] V. I. Levenshtein 等，“能够纠正删除、插入和逆序的二进制编码”，载于《苏联物理学文献》，第10卷，第8期。苏联，1966年，第707–710页。[42] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg 等，“Scikit-learn：Python中的机器学习”，《机器学习研究杂志》，第12卷，第2825–2830页，2011年，出版方：JMLR.org。