融合确定性因子及区域密度的k-最近邻机器翻译方法

齐睿[1,2]，石响宇[1,2]，满志博[1,2]，徐金安[1,2]，陈钰枫[1,2*]
[1]北京交通大学，计算机科学与技术学院，北京，100044
[2]交通数据分析与挖掘北京市重点实验室，北京，100044
{20281284,22120416,zhiboman,jaxu,chenyf}@bjtu.edu.cn

摘要

k-最近邻机器翻译（kNN-MT）是近年来神经机器翻译领域的一个重要研究方向。此类方法可以在不更新机器翻译模型的情况下提高翻译质量，但训练数据中高低频单词的数量不均衡限制了模型效果，且固定的k值无法对处于不同密度分布的数据都产生良好的翻译结果。为此本文提出了一种创新的kNN-MT方法，引入确定性因子（CF）来降低数据不均衡对模型效果的影响，并根据测试点周边数据密度动态选择k值。在多领域德-英翻译数据集上，相比基线实验，本方法在四个领域上翻译效果均有提升，其中三个领域上提升超过1个BLEU，有效提高了神经机器翻译模型的翻译质量。

关键词： 机器翻译；K-最近邻检索；确定性因子

A k-Nearest-Neighbor Machine Translation Method Combining Certainty Factor and Region Density

Rui Qi[1,2],Xiangyu Shi[1,2],Zhibo Man[1,2],Jinan Xu[1,2],Yufeng Chen[1,2*]
[1]Beijing Jiaotong University, School of Computer Science and Technology, Beijing, 100044
[2]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing, 100044
{20281284,22120416,zhiboman,jaxu,chenyf}@bjtu.edu.cn

Abstract

k-Nearest-Neighbor Machine Translation (kNN-MT) is an important research direction in the field of neural machine translation in recent years. It can improve the translation quality without updating the model, but there are still many problems to be solved: the number of each label in training data is not balanced, and the fixed k value can not produce good translation results for the data in different density regions. In this paper, we propose an innovative KNN-MT method to reduce the influence of imbalanced dataset by introducing a certainty factor (CF), and to optimize the search effect by adjusting the k value dynamically according to the data density around the test point. On the multi-domain German-English translation dataset, compared with the baseline experiment, our method has improved the translation performance in all four domains, three of which have improved by more than 1 BLEU.

Keywords: Machine Translation , K-Nearest-Neighbor Search , Certainty Factor

[1]陈钰枫（通讯作者）：chenyf@bjtu.edu.cn

第二十三届中国计算语言学大会论文集，第217页-第229页，太原，中国，2024年7月25日至28日。
卷1：主会论文

1 引言

随着深度学习的飞速发展，神经机器翻译已成为机器翻译任务中的主流方法，翻译效果和质量相比基于统计的机器翻译方法有了巨大飞跃。但目前的机器翻译技术仍然存在着局限性，比如模型的稀有词汇翻译能力不强，在处理不常见或特定领域的术语时，传统的模型往往表现不佳。为此，Khandelwal等人(2021)提出了k-最近邻机器翻译（kNN-MT），kNN-MT是一种基于单词级最近邻检索机制的神经机器翻译增强技术，无需进一步训练，就可以添加到任何预先训练过的神经机器翻译模型中，通过设计数据存储（Datastore）并在其中对相关的术语进行检索，来提高特定领域中单词翻译的准确率，是近年来神经机器翻译中的重要研究方向。

目前，kNN-MT的研究工作主要围绕以下三个方面展开：（1）检索加速化方法：通过减小数据存储体积、优化检索方式等角度提高检索效率。（2）精准检索提升方法：利用预训练模型、监督对比学习等方法对数据存储进行重构，或者调整表示空间使其更加有利于检索。（3）场景特定应用方法：利用kNN-MT模型的优势，提高跨领域、低资源等领域的翻译效果。然而，现有方法主要存在以下两方面问题：

- **数据存储中各分类数量不平衡问题**：传统的kNN算法检索过程中需要考虑到测试点周围各个单词的数量，导致检索生成目标语言单词的概率偏向出现频率高的常见词，使得模型在预测相对罕见单词（rare word）上的表现不佳。**因此，如何提高模型预测罕见单词的能力成为了挑战。**

- **表示向量空间分布不均匀问题**：在检索过程中，现有方法通过定义具体的“k值”来指定检索单词的个数。然而，在数据存储的不同区域内，如果数据密度不同，那么当前进行检索的k值也应当不同，此时的k值应该随着区域密度而产生变化。**因此，如何动态的选择k值成为了另一个需要解决的挑战。**

为了解决以上的挑战，本文提出一种融合确定性因子和区域密度的k-最近邻机器翻译方法（CF-kNN-MT）。具体地，我们引入CF计算模块和动态k值计算模块，前者根据检索到的单词在整个数据存储以及测试点周围区域出现的频率来对kNN概率分布进行调整，后者根据测试点周围区域密度划定一个范围，参考范围内的训练数据来动态确定一个适合测试点的k值。

综上所述，本文贡献总结为如下三点：

- 针对数据存储中高低频单词数量不平衡问题，我们引入确定性因子（CF）来对检索算法进行改进，结合不平衡学习生成新的kNN概率分布，提高了kNN算法修正标准NMT模型预测失误的能力。

- 针对表示向量空间分布不均匀的问题，我们设计了一种根据区域数据密度动态调整k值的算法，以减小各区域密度的差异对检索效果产生的影响。

- 在多领域德-英翻译数据集上，本文提出的模型在各个领域内效果均有提升，其中三个领域上的提升超过1个BLEU，提升较为明显。

2 相关工作

2.1 kNN-MT

k-最近邻机器翻译（kNN-MT）是在标准神经机器翻译的基础上对其结果进行kNN检索以增强其翻译效果的一种方法，其首先将训练数据的上下文表示和目标词作为键值对存储到一个大的数据存储（Datastore）中，然后从数据存储中检索离测试数据最近的k个可能的目标词，以帮助选择输出的下一个单词。

k-最近邻机器翻译(Khandelwal et al., 2021)的流程主要分为两步：数据存储的构建以及机器翻译模型的生成与检索。

构建数据存储：把训练集中源语言的句子作为NMT模型的输入，当输出第t个单词时，以模型的隐藏层表示h_t为键，输出标准答案为值，将构成的键值对储存起来，就构成了数据存储。

翻译生成与检索：将测试集的源语言句子作为NMT模型的输入，将输出的隐藏层表示作为查询向量，在数据存储中检索出离它距离最近的k个键值对，根据其值以及与查询向量的距离来构建一个kNN的概率分布，将其与NMT模型本身的概率分布线性结合在一起，利用构成的综合概率分布对输出的目标进行预测。

目前国内外针对kNN-MT的改进工作主要可以分为三类：研究检索加速化方法、研究精准检索提升方法，以及针对特定场景展开应用。

2.2 检索加速化方法

kNN-MT虽然大幅提高了翻译效果，但其翻译速度远不如标准的NMT模型。因此，许多学者致力于在保持其精度的同时提高kNN-MT的翻译效率，比如基于聚类来对数据存储进行降维和剪枝(Wang et al., 2022a)，既能细化检索结果，又能显著降低推理过程中的翻译延迟。或者为最近邻搜索构建一个更小的数据存储(Meng et al., 2022)，后面解码的步骤中只需要对该数据存储进行检索即可，提高了检索效率。也有学者对检索方式进行了改进，尝试从数据存储中检索整片单词块而不是单个单词(Martins et al., 2022)，或者从输入句子的邻居句子集的子集而不是所有句子中检索邻居目标单词(Deguchi et al., 2023)，使速度得到显著提升。

2.3 精准检索提升方法

同时，也有许多学者通过各种方法进一步提高kNN-MT翻译精度，以使其效果进一步加强。原始的kNN-MT直接使用解码器最后一层的隐状态来构建数据存储，这种方式仍有很大提升空间，因为k-最近邻检索是依据向量之间的相似度进行检索的，而数据存储的向量由标准NMT模型生成，只在机器翻译任务上进行过训练，并没有接受过相似度相关的训练，因此利用其构建的数据存储检索效果并不够优秀。针对该问题，Li等人(2022)提出利用预训练模型构建数据存储，使键相近的值也相近；Wang等人(2022b)则提出利用多个正样本和负样本的监督对比学习来把原来用于检索的隐状态向量h学习成有利于检索的向量z，把数据存储改造成利于检索的形式。而Zhu等人(2023)引入了少量的新参数用来对表示空间进行调整，将密切相关的向量拉进，使表示空间分布平滑，同时使用新的参数异步刷新整个数据存储，以异步获得新的kNN知识。还有一些学者试图对模型本身作出改进：Yang等人(2022)引入k个最近邻知识蒸馏（kNN-KD），训练阶段基础NMT模型直接学习kNN的知识，Jiang等人(2022)通过引入NMT置信度以及进行鲁棒训练来使模型具有具有更好的鲁棒性。

2.4 场景特定应用方法

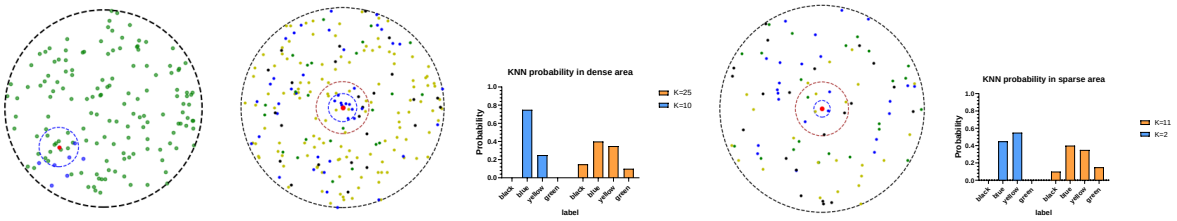
还有一些学者试图将kNN-MT技术应用于特定场景。在跨领域翻译问题上，Cao等人(2023)利用语义相关键-查询对训练出来了一个修订器（reviser），利用其去修改数据存储中键的表示，实现对数据存储的重构，提升模型在下游领域上的翻译效果。而在仅有单语数据的领域内，Huang等人(2023)将kNN-MT引入其中，在正向和反向上使用两个预训练的泛域NMT模型，通过反复迭代实现了NMT的无监督域适应，在对应领域的效果良好。面对低资源领域翻译时，Liu等人(2023)利用kNN-MT进行迁移学习，在初始化、训练和推理全部阶段均转移知识，并选择性地从父数据存储中提取数据构造子感知数据存储（child-aware datastore），有效地降低了模型的过度置信度，同时提高了翻译精度。

以上的三类方法为解决kNN-MT中的核心挑战提供了较好的思路，然而，这些方法仍存在一些局限性：（1）现有方法在提升翻译速度的同时无法保持良好的翻译效果，翻译效率和翻译质量达不到均衡。（2）利用机器翻译语料库构建的数据存储具有不平衡性、不均匀性等问题，均会影响kNN检索的效果。如何使数据存储和kNN检索更好地进行配合，这个问题仍然有待研究。因此，我们进一步提出了融合确定性因子及区域密度的k-最近邻机器翻译方法，主要解决了数据存储具有的不平衡性和不均匀性这两个问题。

3 模型介绍

3.1 问题分析

为了更好的理解本文所解决的问题，我们给出示例进行说明，具体如图1所示。



(a) 数据存储中各分类数量不平衡问题　(b1) 在数据密集区域内进行kNN检索　(b2) 密集区域概率分布示例　(c1) 在数据稀疏区域内进行kNN检索　(c2) 稀疏区域概率分布示例

Figure 1: kNN-MT方法中数据存储存在的问题示例

（1）数据存储中高低频单词数量不平衡问题

由于kNN是基于最邻近的k个样本来决定新样本的类别，如果某一类别的样本数量远多于其他类别，那么新样本更有可能被归为这个多数类别，从而导致分类偏差。我们假设图1a为一个数据存储某一区域的部分训练数据，其中绿色点对应的目标语言单词为A,蓝色点对应的目标单词为B。假设红色点是输入的测试数据，理想情况下应该将其翻译为单词B。我们在数据存储中对其进行kNN检索，可以看到，即使蓝色点基本都聚集在红色点周围，但由于绿色点在整个数据存储中数量都非常多，测试点还是会被分类为绿色点所代表的单词A。对于机器翻译领域，在任何语言中，负责连接和构建句子的功能词（如介词、冠词、连词等）往往出现得更频繁，而负责承载意义的内容词（如名词、动词、形容词等）出现的频率相对较低，这是因为不同的上下文中需要使用不同的内容词来表达特定的意思，但功能词往往可以通用。此外，人们在生活中不断创造新词汇和新表达方式，这些新词相比于传统表达方式出现的频率也更加罕见。因此，在kNN-MT领域针对数据不平衡问题开展研究，对于提高翻译效果至关重要。

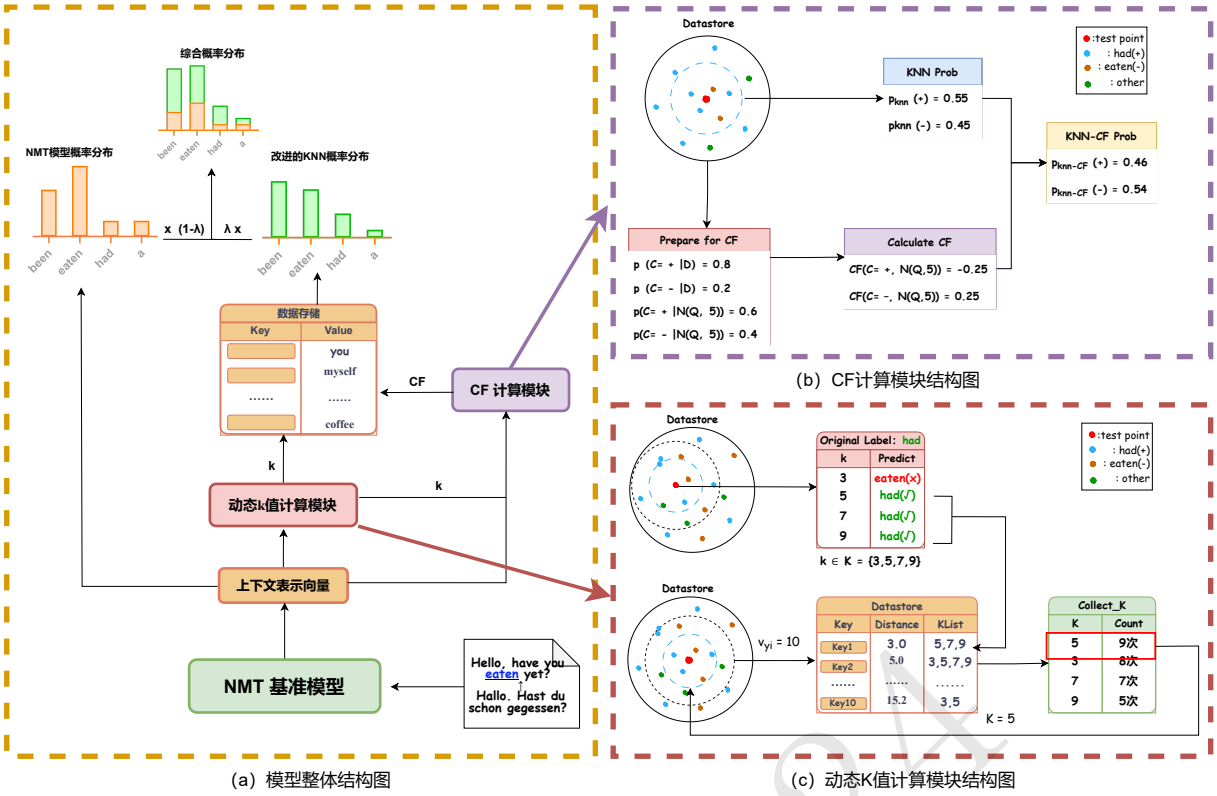
（2）数据存储空间分布不均匀问题

由于语言的多样性和复杂性，数据存储中用于检索的表示向量往往空间分布并不均匀，在不同区域呈现不一致的密度分布，这时使用固定的k值进行检索可能导致不准确的预测结果。假设图1b1和图1c1为一个数据存储的某两个部分，图1b1区域数据比较密集，图1c1区域则比较稀疏。如图1b1和图1b2所示，对于数据密集的区域，最近邻样本大量集中，较小的k值就足以反映出该区域的主要类别特征，较大的k值则容易引入噪声，影响估计的置信度。然而，对于数据稀疏的区域，如图1c1和图1c2所示，即使是最近的向量距离也较远，和测试点相似度不够高。如果取较小的k值（比如k=2），那么该点更有可能被分类为黄色，但由于测试点与检索到的两个最近邻点距离均比较远，特征不够相似，我们不能确信检索结果一定能够代表测试点的真实类别，这时就需要扩大检索的范围，通过检索周围更多数据来增强置信度。因此，我们有必要针对数据存储的空间分布不均匀问题展开研究，对k值进行合理规划。

因此，本文旨在通过解决数据存储中各单词数量不均衡与上下文表示空间密度分布不均匀的问题，提升k-最近邻机器翻译方法的翻译质量。

3.2 模型总体结构

我们对原始的kNN-MT模型进行了改进，整体结构如图2a所示。相比于原始模型，我们添加了CF计算模块，计算时考虑每个单词在整个数据存储中和在测试点周围的出现频率，进而优化kNN检索生成的概率分布；又添加了动态k值计算模块，参考和测试点处于相同密度区域的训练点来为每一个测试数据生成不同的k值，效果强于固定k值的方法。



(a) 模型整体结构图　(b) CF计算模块结构图　(c) 动态K值计算模块结构图

Figure 2: CF-kNN-MT模型结构

3.3 基于确定性因子（CF）测量的kNN分类

3.3.1 确定性因子介绍

确定性因子模型最早由Shortliffe等人(1975)提出，是在基于规则的系统中管理不确定性的一种方法。确定性因子是分布于-1和1之间的数字，用于评估对一个规则的确定性或信心程度，-1表示完全不信任，0表示完全不确定，+1表示完全信任。0到1之间的值表示对命题或假设是真的有不同程度的信心，而0到-1之间的值表示对命题或假设是假的有不同程度的信心。如今，确定性因子在医学诊断、欺诈检测、客户服务、风险分析和自然语言处理等多个人工智能领域有实际应用，并已成功地用于识别数据集(Wu et al., 2004; Zhang and Wu, 2011)中的正关联规则和负关联规则。

3.3.2 CF计算模块介绍

我们在kNN的概率计算中引入了CF，流程如图2b所示：假设D是整个数据存储，里面包含了若干个以上下文表示向量为键，对应单词为值的键值对；Q是测试点的表示向量，$N(Q,k)$是数据存储D中距离Q最近的k个向量所对应值的集合，$p(C=y_i|D)$是数据存储D中值为y_i的键值对数量与所有键值对数量的比值，$p(C=y_i|N(Q,k))$是值为y_i的键值对在集合$N(Q,k)$中的比值。那么我们可以利用上述符号来对CF进行表示。(Zhang, 2010)

如果$p\big(C=y_i|N(Q,k)\big) \geq p(C=y_i|D)$:

$$CF\big(C=y_i,N(Q,k)\big)=\frac{p\big(C=y_i|N(Q,k)\big)-p(C=y_i|D)}{1-p(C=y_i|D)} \tag{1}$$

否则:

$$CF\big(C=y_i,N(Q,k)\big)=\frac{p\big(C=y_i|N(Q,k)\big)-p(C=y_i|D)}{p(C=y_i|D)} \tag{2}$$

我们将CF与kNN概率进行线性组合，其中p_{kNN}表示KNN概率分布，ω表示线性组合的插值参数：

$$p_{kNN-CF}(y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) = \omega\ CF\left(C = y_i, N\left(Q,k\right)\right) + (1-\omega)p_{kNN}(y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) \tag{3}$$

最后将p_{kNN-CF}与原始机器翻译模型生成的概率进行线性组合，就得到了模型输出第i个预测目标的概率分布，其中p_{MT}表示原始机器翻译模型的概率分布，λ表示线性组合的插值参数：

$$\mathrm{p}\ (y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) = \lambda\ p_{kNN-CF}(y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) + (1-\lambda)p_{\mathrm{MT}}\ (y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) \tag{4}$$

计算CF值的伪代码如代码1所示。

Algorithm 1 Calculate CF

Require:
y_i：模型预测的第i个单词
K_{num}：在测试点周围选取的邻居数量（K）
K_{dic}：测试点周围K个邻居所对应单词的集合
D_{num}：整个数据存储的键值对数量
D_{dic}：整个数据存储中单词的集合
function CalculateProb:(y_i, num, dic):
  $count = dic$集合中元素y_i的数量
  $p = count/num$
  return p

  $p1$ = CalculateProb(y_i, K_{dic}, K_{num})
  $p2$ =CalculateProb(y_i, D_{dic}, D_{num})
  if $p1 \geq p2$ **then**
    $CF = (p1 - p2)/(1 - p2)$
  else
    $CF = (p1 - p2)/p2$
  end if
  return CF

3.4 基于区域密度表示的动态k值算法

3.4.1 局部密度与最近邻的关系

定理：(Mullick et al., 2018)假设一个d维超球面包含一个位于中心的点和内部一些均匀分布的点。当超球面包含的点数为N_1和N_2时，设距离中心最近的点与中心的期望距离分别为d_{N1}和d_{N2}。如果N_1>N_2 ，则d_{N1}<d_{N2}。

从定理可以看出，局部数据密度与最近邻距离近似成反比。因此，如果某一局部中某一数据的最近邻距离较大，我们可以假设该数据所在区域数据较为稀疏，而反之则表示区域数据比较密集。因此，我们使用最近邻距离作为查询点所在区域的密度表示，并按比例改变kNN-MT选取邻域的大小。

3.4.2 动态k值计算模块介绍

为减小数据存储中各个区域的数据密度不均匀对结果造成的影响，我们设计了一种动态k值选取方法，尝试为每个查询点估计一个合适的k值，从而用动态的k值进行最近邻检索。

我们算法的核心思想如图2c所示：在预处理阶段将数据存储中的每一个键值对看做测试点，分别选取不同的k值，将该点在除本身以外的数据存储中进行检索，把能够正确分类该点的k值记录下来,形成集合KList,把每个键值对和对应的KList集合一同存储进数据存储中。在对测试数据进行检索时，我们参考每个测试点周围的v_{y_i}个邻居所适合的k值，选定检索该测试点所用的k值。确定k值的流程如下：计算CF值的伪代码如代码1所示。

Algorithm 2 Build KList

Require:
K=[3,5,7,9]:备选K集合
KList[key]:能够将键值对(key,value)正确分类的k值构成的集合
D:数据存储，里面存有许多个键值对(key,value)

```
for (key,value) in D do
  for k' in K do
    if 对key进行kNN检索(k=k')，检索出来的结果为value then
      k' -> KList[key]
    end if
  end for
end for
return KList
```

（1）确定v_{y_i}:

对于数据存储中任意一个键值对，我们假设键为x_i,我们将除它本身之外距离他最近的键值对与其的距离设为$d_{1NN}^{P/x_i}(x_i)$,即为最近邻距离。遍历数据存储中的所有键值对，计算每一个键值对的最近邻距离，其中最大的最近邻距离记为d_{max}，最小的最近邻距离记为d_{min}。即:

$$d_{max} = \max_{x_i \in P} d_{1NN}^{P \setminus \{x_i\}}(x_i) \tag{5}$$

$$d_{min} = \min_{x_i \in P} d_{1NN}^{P \setminus \{x_i\}}(x_i) \tag{6}$$

接下来，我们分三种情况讨论v_{y_i}:

$$v_{y_i} = \begin{cases} v_{\max}, & d_{1NN}^{P}(y_i) < d_{\min} \\ 1, & d_{1NN}^{P}(y_i) > d_{\max} \\ \left\lceil \left(v_{\text{lin}}\left(y_i\right) v_{\exp}\left(y_i\right)\right)^{0.5} / \delta \right\rceil, & d_{\min} \le d_{1NN}^{P}(y_i) \le d_{\max} \end{cases} \tag{7}$$

其中$\lceil . \rceil$符号代表向上取整, δ 是负责调整v_{y_i}大小的参数,$v_{\max} = \sqrt[3]{n}$。也就是说，若y_i周边数据足够密集，v_{y_i}被定为v_{max}；若y_i周边数据过于稀疏，v_{y_i}被定为1。在普遍情况下，$v_{y_i} \in [1,\ v_{max}]$。

其中，

$$v_{\text{lin}}(y_i) = [\frac{(1-\nu_{max})}{(d_{\max} - d_{\min})}(d_{1NN}^{P}(y_i) - d_{\min}) + \nu_{max}] \tag{8}$$

$$v_{exp}(y_i) = v_{max} e^{(d_{1nn}^{p}\ (y_i) - d_{min}\ )\beta} \tag{9}$$

$$\beta = -\frac{log_e v_{max}}{d_{max} - d_{min}} \tag{10}$$

在上述公式中，线性估计$v_{lin}\left(y_i\right)$将v_{y_i}建模为$d_{1NN}^{P}\left(yi\right)$在$v_{max}$和1之间的一个线性递减函数，对于较小的数据集能够维持一个可观的邻域大小；而指数估计$v_{exp}(y_i)$则是$d_{1NN}^{P}\left(y_i\right)$的指数递减函数，可以在较大的数据集中限制邻域不要过大。接下来，我们取这两种估计的几何平均值，因为相较于算术平均值而言，几何平均值更接近两个估计中较小的一个（即指数估计），该均值允许指数估计比线性指数估计对v_{y_i}的计算有更大的影响，这样的话即使对于较大的数据集，影响测试点的k值的邻域也可以保持较小。

（2）k值选择:

对于我们要进行检索的测试点y_i,假设上式计算出的$v_{y_i} = v$。对于一组函数$F = \{ f_1,\ f_2, \ldots\ f_{kmax} \}$，每一个$f_k \in F$ 都是从数据集X到{0,1}集合的多对一映射。其中$k_1, k_2 ..... k_{max} \in K$,$K$为可能会被用来进行kNN检索的k值的集合。

$$f_{k'}(x) = \begin{cases} 1\text{,如果}x \in X\text{且当}k = k'\text{时，}x\text{被}kNN\text{正确分类} \\ 0\text{,除此之外的情况} \end{cases} \tag{11}$$

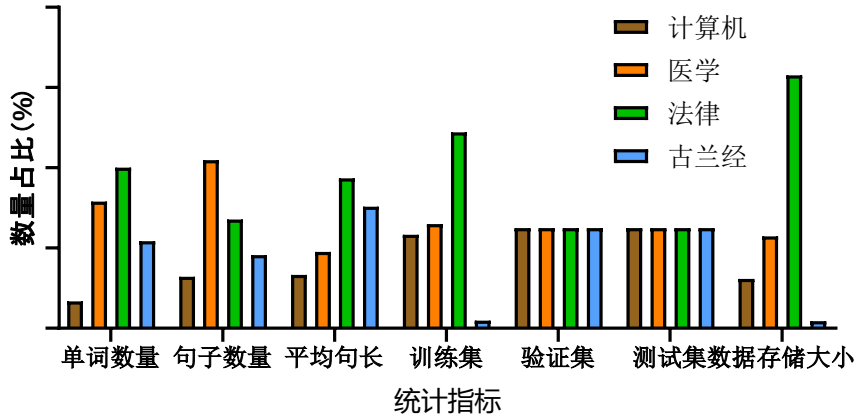


Figure 3: 不同领域数据集信息统计图

接下来我们定义:

$$z_{k'} = \sum_{x_i \in \tau_v^P(y_i)} f_{k'}(x_i) \tag{12}$$

$z_{k'}$表示当选取的k值为k'时，把数据存储中在y_i周围的每个键值对当做测试点，在数据存储中进行搜索，有多少个键值对可以被正确分类。因此，继续定义：

$$k_{yi} = \underset{k' \in \{1,2,..,k_{max}\}}{\text{argmax}} \quad z_{k'} \tag{13}$$

取得的k_{yi}即为用来对测试点y_i进行检索的k值。

4 评测实验与分析

4.1 实验设置

4.1.1 实验数据集

为了验证该模型的效果，我们在几个常用的数据集上进行了实验：我们使用多领域德-英翻译数据集(Koehn and Knowles, 2017)进行德语-英语语言对实验，该数据集被广泛使用于机器翻译模型的效果评测中。我们跟随原始kNN-MT模型，在四个常用的基准上进行实验，其中的领域包括计算机、古兰经、医学和法律，以上数据集的细节如表1 所示。

数据集	计算机	医学	古兰经	法律
单词数量	3,041,677	14,301,472	9,848,539	18,128,173
句子数量	337,817	1,104,752	480,421	715,372
平均句子长度	9.0	12.9	20.5	25.3
训练集	223K	248K	18K	467K
验证集	2K	2K	2K	2K
测试集	2K	2K	2K	2K
数据存储大小	3.71M	6.90M	524K	19.0M

Table 1: 不同领域数据集信息及对应数据存储大小

我们对四个领域的数据集进行对比分析，从图3中我们可以分析得出，计算机领域数据集单词量较少，句子平均长度较短，反映了技术文档的直接性和命令式的表达方式；法律领域单词量多，且平均句长最长，显示了法律文本的复杂性和条款的详尽性。而医学领域句子数量最多，且长度较短，反映出医学文本对结构性和准确性的需求；古兰经领域的句子长度较长，文本特征更为复杂，反映了宗教文本的语言风格，增加了翻译的复杂性和挑战性。

4.1.2 对比模型

Base NMT：我们的基准模型，是一个基于大型Transformer架构的编码器-解码器模型，使用Fairseq工具包实现。模型使用了WMT'19德英新闻翻译任务的双语数据以及反向翻译和人工翻译数据进行训练，然后在newstest数据集上进行微调。在数据处理阶段，作者使用Moses标记器对所有数据进行标准化和标记化处理，使用32K联合字节对编码（BPE）进行子词分割。该模型可以在fairseq的库中获取到，原论文及模型下载链接：https://github.com/facebookresearch/fairseq/tree/main/examples/translation。

Vanilla kNN-MT：我们的基线模型(Khandelwal et al., 2021)，本文工作主要是在此模型的基础上进行改进。模型利用多领域德-英数据集为每个领域专门构建数据存储，翻译过程中将Base-NMT模型的输出结果在域内数据构建的数据存储中进行kNN检索以增强其翻译效果。数据处理阶段同样需要对数据进行标准化和BPE（字节对编码）处理。模型具体流程介绍见2.1节，模型代码下载链接：https://github.com/urvashik/knnlm。

N-Gram kNN-MT：我们和Lv等人(2023)的模型结果进行比较，该方法在原始kNN-MT模型的基础上进行改进，将相邻的n-gram表示向量连接起来作为键，而对应目标单词的元组作为值。

4.1.3 超参数及实验设置

本着控制变量原则，在会影响模型训练结果的超参数上，我们尽量保持统一。

温度参数T：用以防止模型对单个近邻给予过高的概率，进而防止模型过拟合到最相似的近邻点。大于1的温度参数会使分布变平滑，从而提高kNN-MT的性能。在古兰经领域我们取T = 100，其他领域均取T = 10。

插值参数λ：用以控制kNN检索概率和NMT预测概率线性组合的权重，在计算机领域我们取$\lambda = 0.7$，其他领域均取$\lambda = 0.8$。

波数beam：用以控制波束搜索的束的个数，我们取beam = 4。

上述参数我们都与原始kNN-MT模型保持一致。此外，本文模型中新加入了两个参数：

调节参数δ：用于控制v_{y_i}的大小，它与数据存储的大小有关。

新插值参数ω：用于将CF值与kNN概率进行线性组合。

K集合：在预处理过程中用于对数据存储每个点进行检索的k值的集合，后续对测试点进行检索时使用的是该集合中的其中一个k值。

新加入参数的具体数值如表2所示。

数据集	δ	ω	K集合
医学	0.25	0.5	3,5,7,9,11
IT	0.25	0.6	3,5,7,8,9,11,13,15
法律	0.15	0.8	3,5,7,8,9,11
古兰经	0.25	0.4	7,11,15

Table 2: 本实验有关超参数设置

4.1.4 评测指标

在本文实验中，所有的翻译结果都由SacreBLEU(Post, 2018)用区分大小写的去标记化BLEU进行评估。BLEU的主要原理是通过比较候选译文与参考译文中的n-gram的匹配程度来评估译文质量，匹配程度越高，我们认为译文的质量越优。BLEU评分得分范围从0到1，1表示完美的匹配。SacreBLEU是一个用于计算机器翻译评估指标BLEU分数的常用工具,它采用WMT（机器翻译会议）标准标记法，可以正确计算去标记化输出的分数.

BLEU可能无法完全捕捉翻译的流畅性和语言多样性，并且对于不同的参考翻译可能存在一定的敏感性，因此我们还采用COMET(Rei et al., 2020)指标作为补充度量。与传统的基于n-gram匹配的BLEU评分不同，COMET利用新的跨语言预训练模型（如XLM-RoBERTa），通过比较机器翻译输出、参考翻译和源文本的语义相似性来评估翻译质量，能够更全面地反映翻译的准确性、流畅性和忠实度。

4.1.5 模型细节

我们以Zhu等人(2024)研发的kNN-BOX框架[0]为基础，使用Faiss[1]来进行向量检索。本文实验均在一个NVIDIA 2080 Ti GPU上进行。

4.2 实验结果与分析

4.2.1 实验结果

我们分别将各个领域的训练集数据构建成数据存储，再用相应领域的测试集进行测试，结果如表4所示。表中标记为"-"的数据表明原作者未在文章中给出。

模型	计算机	法律	医学	古兰经	均值
Base NMT	38.35/39.21	45.48/57.52	40.06/46.93	16.26/-1.37	35.04/35.57
Vanilla kNN-MT	45.58/49.64	61.01/66.03	54.19/53.46	20.38 /-0.53	45.37/42.15
N-Gram kNN-MT	46.42/-	61.98/-	55.27/-	**21.39**/-	**46.26**/-
本文模型	**46.73/50.35**	**62.15/66.12**	**55.29/54.84**	20.86/**0.94**	**46.26/43.06**
模型提升	+1.15/+0.71	+1.14/+0.09	+1.10/+1.38	+0.48/+1.47	+0.97/+0.91

Table 3: 在不同领域的德-英翻译测试集上的评测得分（BLEU/COMET）

从表中可以看出，与之前的研究相同(Khandelwal et al., 2021)，原始kNN-MT在所有数据集上的表现都显著优于基础NMT。此外，本文模型在全部领域的BLEU评分相较基线系统均有明显提升，BLEU分数平均值超过原始的kNN-MT达到了0.97点。而在计算机、法律、医学三个领域，本文模型提升超过1个BLEU，且提升效果优于对比模型N-Gram kNN-MT。此外，本文模型在所有领域的COMET评测得分也高于基准模型。综合考虑两个评测指标，我们发现本文模型在医学领域的提升效果最为明显，根据表1我们猜测这是因为医学领域单词数量较多，里面包含许多专业术语，这些术语属于低频词汇，我们的CF模块可以让翻译模型加强对这些低频词汇的关注。

4.2.2 案例分析

我们在四个领域的测试数据中选取了一些典型的例子，分别输出原始模型和本文模型的翻译结果，进行对比分析，如图4所示。我们可以看出，本文模型的翻译结果相比原始模型更加贴近标准答案，在计算机和法律这两个领域的翻译结果更加精准，用词更加专业；在更要求严谨性的医学领域，我们的表述更加准确，且表达方式与源语言更加相似，不易产生歧义。古兰经领域数据集中包含着具有高度的专业性和语境依赖性的宗教文本，现有模型翻译效果均不够优秀，但我们的模型在用词的准确性和句意的表现上更胜一筹。

4.2.3 消融实验

为了验证CF计算模块和动态k值计算模块对模型性能的贡献，我们进行了一系列消融实验，结果如表4所示。为了保证结果的可比性，除了被移除的模块，其他所有设置均保持不变。

从结果中我们可以看出，分别移除CF模块和动态k值计算模块，以及同时移除两个模块之后，模型的性能相比完整模型在各个领域数据集上的BLEU值均有降低，这表明上述两个模块对翻译准确性的提高均起到了重要作用。

4.2.4 探究实验

为了使实验结果更加可靠，我们进一步研究了一些重要的超参数对模型的影响，以及一些变量的取值。

关于v_{y_i}的取值 在计算v_{y_i}的取值时，我们要考虑y_i周边区域的局部信息，比如数据密度以及邻域数据的类型分布，把它限制在一个合理的范围。如果周边数据较为密集，那么我们可以选取较大的v_{y_i}，这样选择k值的时候可以参考更多周边邻居的意见；反之如果周边数据较为

[0] https://github.com/NJUNLP/knn-box
[1] https://github.com/facebookresearch/faiss

计算机	医学
[源语言]:Beschreibe die Merkmale der einzelnen Stufen [标准答案]:Describe the **level characteristics** [KNN-MT]:Describe the **properties of each level** [CF-KNN-MT]:Describe the **level characteristics** [源语言]:zur Auswahl einer Hintergrundfarbe für die Präsentation. [标准答案]:**dialogue box, with which you can** select the background colour **of** the presentation. [KNN-MT]:**to** select a background color **for** your presentation. [CF-KNN-MT]:**dialogue box, with which you can** select a background color **for** the presentation.	[源语言]:Atmen Sie so weit wie möglich aus. [标准答案]:Breathe out as far as **is comfortable**. [KNN-MT]:Breathe out as far as **it will go**. [CF-KNN-MT]:Breathe out as far as **is comfortable**. [源语言]:Anwendung bei Kindern und Jugendlichen unter 18 Jahren. [标准答案]:Use in children and adolescents **under the age of 18 years.** [KNN-MT]:Use in children and adolescents **below 18 years of age.** [CF-KNN-MT]:Use in children and adolescents **under the age of 18 years.**
法律	**古兰经**
[源语言]:Die Aufstellung muß ausserdem folgende Angaben enthalten [标准答案]:The recapitulative statement **shall also set out** [KNN-MT]:The recapitulative statement **must also include the following information:** [CF-KNN-MT]:The recapitulative statement **shall also set out** [源语言]:Keine weiteren Berichtigungen wurden beantragt oder für notwendig gehalten. [标准答案]:No **further adjustment** was **claimed** or **considered** necessary. [KNN-MT]:No further **corrections** were **requested** or **deemed** necessary. [CF-KNN-MT]:No **additional** adjustments were **claimed** or **considered** necessary.	[源语言]:Meine Macht hat mich verlassen. [标准答案]:Vanished has my power from me.（意译：**我的力量消失了**） [KNN-MT]:I am not able to escape.（意译：**我无路可逃**） [CF-KNN-MT]:My power has left me.（意译：**我的力量消失了**） [源语言]:Gott hat sich Abraham ja zum Vertrauten genommen. [标准答案]:**And** God **chose** Abraham as friend. [KNN-MT]:**Verily,** God **has taken** Ibrahim as friend. [CF-KNN-MT]:God **chose** Abraham as a friend.

Figure 4: 对原始模型和本文模型翻译效果进行对比分析

模型	计算机	法律	医学	古兰经	均值
CF-kNN-MT	46.73	62.15	55.29	20.86	46.26
-CF	45.97	61.22	54.40	20.59	45.56
- 动态k值	45.75	61.86	54.66	20.54	45.70
-（CF+动态k值）	45.88	61.01	54.19	20.38	45.37

Table 4: 在德-英翻译数据集上进行消融实验的BLEU得分

稀疏，那我们应该选取较小的v_{y_i}，以剔除掉距离过远的邻居的影响。在实验中，我们使用公式7来计算v_{y_i}的取值，与上述理论保持一致。

关于K集合的选取 在实验之初，我们选取了较为广泛使用的1,3,5,7,9作为备选k值(Mullick et al., 2018)，对其进行多种排列组合，将不同组合放入K集合中，在多个数据集上进行测试，发现K集合中的值不可以过少，因为这样会限制最后对测试数据检索时可供选择的k值，使动态k值的效果不够明显，极端情况下就变成了固定k值的检索。但备选k值并非越多越好，因为这可能会增加模型在决策时的不确定性，特别是某些较大的k值可能并不适合给定的数据分布，还会增加预处理的时间与占用的空间。根据我们在四个基准数据集上的评测结果，每个数据集上的K集合已经在表2中说明。

这里我们注意到一个问题：如果K集合中的某些k值能够正确分类的周边训练点数量一样且均为最多，那该选择谁作为检索测试点的k值呢？在这里，我们有六种策略：取最小的k值（方法A）、最大的k值（方法B）、随机选取一个k值（方法C），选择均值（方法D），选择中位数（方法E），以及和距离测试点最近的键值对选择同样的k值（方法F）。我们以古兰经数据集为例，针对四种策略进行了对照实验，结果如表5所示。表中结果表明，方法F的效果最为良好，且结果稳定，我们分析这是因为k值的选择受周围数据密度影响很大，而距离测试点最近的点和测试点周边的密度最为相似，因此其对k值的选择和测试点的选择也更为相似，所以要侧重考虑。

探究确定性因子的影响 在本文中，确定性因子被用来表示某一类别在整个数据存储的某一部分中出现的频率与在整个数据存储中出现的频率之比。$CF\big(C=y_i|N(Q,k)\big)$的取值为[-1,1]，如果CF大于0，那么我们认为将测试点的值预测为y_i的概率应该增加；CF小于0，我们认为预测为y_i的概率应该减小。CF等于0，我们认为目标点应该被预测为$C=y_i$的概率与正常kNN相同。在探究确定性因子要如何影响kNN检索结果时，通过反复比对，我们发现令kNN概率与确定性因子线性结合能够得到最好的效果。

探究动态k值的影响 如3.1中所描述的那样，当测试点位于密度不同的区域时，能够使检索

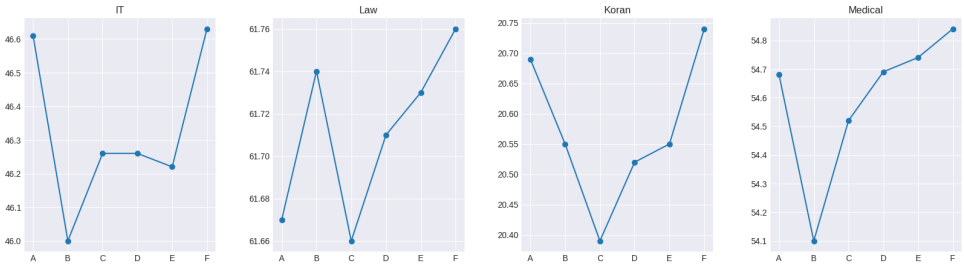


Figure 5: 面对多个k值均为最优时，不同方法选择对于模型效果的影响

效果达到最好的k值也不同，因此我们有必要根据测试点周围的密度动态选取k值。而在本文的方法中，k值的选择要参考测试点周围的数据，因为测试点周围的数据和测试点所处同样一片区域，周围的数据密度几乎一样，如果某一k值能够将测试点周边数据都正确分类，那么它也适合对测试点进行分类。

5 总结

本文针对k-最近邻机器翻译研究中存在的数据存储中高低频单词数量不平衡以及上下文表示空间密度分布不均匀两个尚未被关注的问题，提出了一种改进的神经机器翻译增强算法。在此过程中，我们取得了以下成果：

1.针对低频词较难被检索算法关注的问题，我们引入了确定性因子来增加这些数据被检索到的概率。确定性因子的引入在一定程度上平衡了各单词的数量不均衡问题，从而提高了kNN算法修正标准NMT模型预测失误的能力。

2.针对数据存储空间上密度分布不均匀的情况，我们设计了一种根据区域密度动态选取kNN检索所需的k值的算法。通过动态调整k值，我们能够减小不同区域数据密度差异对检索结果的影响，从而提高了检索的精度和可靠性。

在多领域德-英翻译数据集上，本文提出的改进神经机器翻译增强算法在各个领域内效果提升明显，均高于基准模型。未来我们将针对领域适应性开展研究，对该模型进行进一步改进，提升其在跨领域翻译等方面的性能。

致谢

本研究受国家自然科学基金面上项目(No.62376019,61976015, 61976016,61876198,61370130)资助。作者们还对匿名评审专家给予的宝贵建议表示衷心的感谢。

参考文献

Zhiwei Cao, Baosong Yang, Huan Lin, Suhang Wu, Xiangpeng Wei, Dayiheng Liu, Jun Xie, Min Zhang, and Jinsong Su. 2023. Bridging the domain gaps in context representations for k-nearest neighbor neural machine translation. pages 5841–5853. ACL.

Hiroyuki Deguchi, Taro Watanabe, Yusuke Matsui, Masao Utiyama, Hideki Tanaka, and Eiichiro Sumita. 2023. Subset retrieval nearest neighbor machine translation. pages 174–189.

Hui Huang, Shuangzhi Wu, Xinnian Liang, Zefan Zhou, Muyun Yang, and Tiejun Zhao. 2023. Iterative nearest neighbour machine translation for unsupervised domain adaptation. pages 13294–13301.

Hui Jiang, Ziyao Lu, Fandong Meng, Chulun Zhou, Jie Zhou, Degen Huang, and Jinsong Su. 2022. Towards robust k-nearest-neighbor machine translation. pages 5468–5477.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Jiahuan Li, Shanbo Cheng, Zewei Sun, Mingxuan Wang, and Shujian Huang. 2022. Better datastore, better translation: Generating datastores from pre-trained models for nearest neural machine translation. Number 1. Cornell University Library, arXiv.org, 2022-01-01. Report ER -.

Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao, and Min Zhang. 2023. knn-tl: k-nearest-neighbor transfer learning for low-resource neural machine translation. pages 1878–1891. ACL.

Rui Lv, Junliang Guo, Rui Wang, Xu Tan, Qi Liu, and Tao Qin. 2023. N-gram nearest neighbor machine translation, 01.

Pedro Henrique Martins, Zita Marinho, and Andr E. F. T. Martins. 2022. Chunk-based nearest neighbor machine translation. pages 4228–4245. EMNLP.

Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. 2022. Fast nearest neighbor machine translation. pages 555–565. ACL.

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. 2018. Adaptive learning-based k -nearest neighbor classifiers with resilience to class imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5713–5725.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Edward H. Shortliffe and Bruce G. Buchanan. 1975. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3):351–379.

Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. 2022a. Efficient cluster-based k-nearest-neighbor machine translation. pages 2175–2187. ACL.

Qiang Wang, Rongxiang Weng, and Ming Chen. 2022b. Learning decoupled retrieval representation for nearest neighbour neural machine translation.

Xindong Wu, Chengqi Zhang, and Shichao Zhang. 2004. Efficient mining of both positive and negative association rules. *ACM Trans. Inf. Syst.*, 22(3):381–405, jul.

Zhixian Yang, Renliang Sun, and Xiaojun Wan. 2022. Nearest neighbor knowledge distillation for neural machine translation. pages 5546–5556, July. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Carpuat, Marine de Marneffe, Marie-Catherine Meza Ruiz, Ivan Vladimir.

Shichao Zhang and Xindong Wu. 2011. Fundamentals of association rules in data mining and knowledge discovery. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 1:97–116, 03.

Shizhao Zhang. 2010. Knn-cf approach: Incorporating certainty factor to knn classification. *IEEE Intell. Informatics Bull.*, 11(1):24–33.

Wenhao Zhu, Jingjing Xu, Shujian Huang, Lingpeng Kong, and Jiajun Chen. 2023. Ink: Injecting knn knowledge in nearest neighbor machine translation. pages 15948–15959. ACL.

Wenhao Zhu, Qianfeng Zhao, Yunzhe Lv, Shujian Huang, Siheng Zhao, Sizhe Liu, and Jiajun Chen. 2024. knn-box: A unified framework for nearest neighbor generation. In Nikolaos Aletras and Orphée De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*, pages 10–17. Association for Computational Linguistics.