arXiv:2304.01352v2 [cs.CL] 5 Apr 2023

一种简单有效的跨语言抄袭检测方法

Karen Avetisyan[a, b, *], Arthur Malajyan[a, b], Tsolak Ghukasyan[a, b], Arutyun Avetisyan[a, c]

[a]*Institute for System Programming of the Russian Academy of Sciences, Alexander Solzhenitsyn, 25, Moscow, Russia, 109004*[b]*Russian-Armenian University, Hovsep Emin, 123, Yerevan, Armenia, 0051*[c]*ISP RAS Research Center for Trusted Artificial Intelligence, Alexander Solzhenitsyn, 25,Moscow, Russia, 109004*

摘要

我们提出了一种简单的跨语言抄袭检测方法，适用于大量语言。该方法利用开放的多语言词库执行候选检索任务，并使用预训练的多语言 BERT 语言模型进行详细分析。该方法在使用时不依赖机器翻译和词义消歧，因此适用于包括资源匮乏语言在内的大量语言。我们在若干现有基准和新基准上展示了所提方法的有效性，并在法语、俄语和亚美尼亚语上取得了最先进的结果。

关键词：跨语言抄袭检测、基于词库的抄袭检测、资源匮乏语言、多语言词簇

1. 引言

抄袭检测自动化是确保学术界及生活其他领域诚信与公平晋升的重要组成部分。

*通讯作者*Email addresses:* karavet@ispras.ru (KarenAvetisyan), malajyanarthur@ispras.ru(Arthur Malajyan), tsggukasyan@ispras.ru (Tsolak Ghukasyan), arut@ispras.ru (Arutyun Avetisyan)

单语抄袭检测已取得显著进展，即检测同一种语言的文本之间是否存在抄袭。许多研究也致力于解决跨语言抄袭检测问题[1, 2, 3]。与此同时，正如 [4]所强调的，针对资源匮乏语言的研究和资源明显不足。

存在多种方法可以克服跨语言抄袭检测中的语言差异。将所有文档转换为同一种语言是最直观的解决方案，而且已在许多研究中成功应用 [5, 6, 7, 8, 9]。然而，这种方案依赖高质量的机器翻译系统，而许多语言对并不具备这样的系统。此外，翻译整篇文档既耗时又昂贵。还有一些基于语料库的方法，利用不同语言的平行文本，训练将文档映射到共享空间中的向量的方法 [10, 11]。由于缺乏相关语料，这些方法也不适用于大多数语言对。基于词库的方法则提供了机器翻译和基于语料库的方法之外的替代方案，例如使用跨语言词嵌入和多语言语义网络来克服语言障碍。本文采用了类似的方法。

整合多语言词库中的知识是这些方法的基础组成部分。 [12] 表明，词义消歧在这方面非常有效。然而，消歧任务可能颇具挑战性，尤其是对于词汇丰富且资源匮乏的语言。另一种方法是采用模糊搜索，如 [13]所示。该方法假设，如果文本上下文足够充分，真正的概念就会具有显著更高的概念质量，从而减少错误；但遗憾的是，情况并非总是如此，尤其是对于词汇丰富的语言。因此，在这项工作中，我们提出一种替代方法：根据简单的频率近似来合并词库，而非使用词义消歧。所提出的抄袭检测方法在法语和俄语基准测试上取得了最先进的结果，并且也证明适用于亚美尼亚语等资源匮乏语言。

在详细分析阶段，我们使用预训练多语言掩码语言模型比较不同文本的原文段落，以确定其语义相似度。这一思路源于基于 BERT 的语言模型在释义检测任务 [14, 15], 中的成功；该任务在本质上与目标任务非常相似，同时也得益于这些模型的多语言版本 [16]的出现。

本研究的主要贡献如下：

1. 一种高效、准确的跨语言抄袭检测方法，该方法在运行时无需使用机器翻译，也不依赖词义消歧。2. 一种基于多语言叙词表构建多语言词簇的方法，用于资源匮乏语言的跨语言信息检索，对大语种也具有竞争力。3. 一个新的跨语言抄袭检测测试数据集，涵盖 5 种语言对。4. 一个具有挑战性的句子级跨语言语义相似度分析测试数据集，涵盖 5 种语言对。

本文结构如下：第 2 节回顾相关工作，第 3 节介绍所提出的方法，第 4 节给出性能评估结果，并在第 5 节进行讨论。

我们在 GitHub 上发布了所有展示的数据集和模型。[1]

2. 相关工作

本节介绍此前用于跨语言抄袭检测任务的一些方法。与单语系统类似，跨语言场景中的方法也主要采用两阶段方法 [17]。第一阶段是从大型参考语料库中检索少量可能相关的文档。其重点是缩小第二阶段的搜索范围

[1] https://github.com/1998karen/Cross-Lingual-Plagiarism-Detection

；第二阶段则着重对可疑文档和检索到的文档进行更详细的比较，以提高精确率。第一阶段处理的文档可以是全文，也可以是较小的片段，例如段落、句子或固定大小窗口。第二阶段即详细分析，会将可疑文档与上一步检索到的源文档逐一比较。此阶段的目标是对齐经过分析的文档中相似的文本部分。本文提出的方法同样采用两阶段方法。

基于句法和词汇的方法 在计算相似度时不使用任何多语言数据或翻译机制，主要用于句法和词汇彼此相近的语言。例如，CL-CNG[18]就是这类方法之一，它使用字符 n 元语法向量来表示文本。

基于词库的方法 通过将文本转换为语言无关的形式，再对转换后的文本进行比较。BabelNet [19] 或 EuroWordNet [20] 等词库是由不同文本单元构成的多语言语义网络。对于每个词，MLPlag [21]使用 EuroWordNet 叙词表提取语言无关的词义，随后直接对其进行比较。基于跨语言概念词库的相似度（CL-CTS）方法 [22] 使用 Eurovoc[2] 概念词库将文档转换为向量，然后通过余弦相似度比较这些向量。另一种方法是基于知识图谱的 CL-KGA 方法 [23, 24] ，该方法使用 BabelNet。它会根据词语的概念及其间的路径，为每个文本片段构建图。最后，使用某种图相似度算法计算相似度。所述方法通常依赖词义消歧，因此无法用于缺少此类工具的资源匮乏语言。

基于语料库的方法使用从多语言数据库中提取的不同信息。跨语言显式语义分析（CL-ESA）模型 [10]是显式语义分析 [25]的多语言版本。假设

[2] https://op.europa.eu/s/vFSH

D 是一个文档集合，对于该集合中的每个文档d ，ESA算法都会为其关联一个概念向量 $\mathbf{d}$，而该向量是通过计算所研究文档与维基百科中每个文章概念之间的 TF-IDF相似度构建的。CL-ESA方法的工作原理完全相同，但它使用的是概念对齐的多语言语料集合，而不是单语语料集合 D。另一种基于平行语料库的方法是跨语言对齐相似度分析（CL-ASA） [11], ，它使用通过 IBM 对齐模型 1 [26]获得的统计双语词典。这些方法的缺点是需要训练集，而训练集仅适用于少数语言对。

基于机器翻译的方法 将文档翻译成同一种语言，然后使用单语搜索算法。许多作者采用了基于翻译加单语（T+MA）的方法 [5, 6, 7, 8, 9]。与基于语料库的方法类似，对于大多数语言对而言，使用机器翻译并不可行，因为开发准确的翻译器或获取其使用权限的成本极高。

基于词嵌入的算法。 这些方法利用词语的向量表示来比较文本。文献 [13] 提出了两种方法：CL-CTS-WE 使用每个词最接近的 10 个词嵌入，为该词构建词袋。第二种方法 CL-WES 则将每个待考察的文本单元表示为其所含词语的向量之和。这两种方法所用向量的双语特性由 Multivec 提供 [27]。另一种基于词嵌入的方法 [28]使用了来自 [16] 的双语词向量。

除了上述单独的方法外， [29] 或[13] 等其他方法还尝试融合这些方法，以取得更好的效果。

所提方法属于基于词库的方法，但与大多数方法不同，它不依赖词义消歧。据我们所知，所提方法是首个将 Universal WordNet 用于多语言抄袭检测任务的方法。

3. 所提方法

本节介绍用于跨语言抄袭检测的两步法。下文的小节将说明如何从大量参考文档中检索出少量相关文本段落。之后，我们将介绍详细分析步骤，在该步骤中，使用更精确的方法将检索到的段落与可疑段落进行比较。

下文将使用“可疑文档/片段”和“参考文档/片段”这两个术语。可疑文档/片段指可能包含抄袭内容的文档/片段。参考文档/片段指可能被用来进行抄袭的文档/片段。

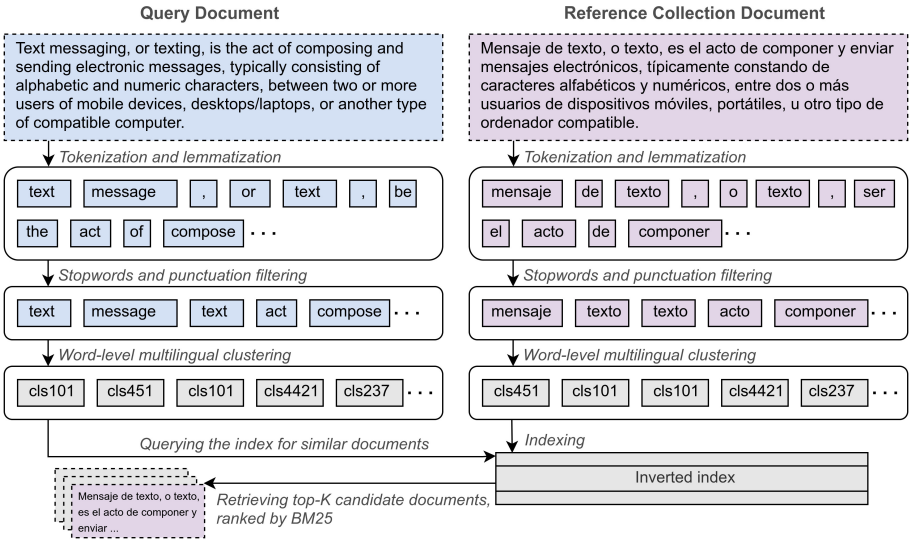

图 1：查询文档与参考文档集预处理、词级多语言聚类和相关片段检索方案。

3.1. 候选检索

候选检索过程是一个相关文档提取过程。它减少了需要经过计算成本高昂的详细分析步骤的文档数量。在所提方法中，不使用全文

而是将可疑文档和参考文档都划分为较小的文本片段（例如段落），并在该层级进行相关候选检索。

我们将任务定义如下：给定一组使用语言 L_1 的文本段落 D 、一段使用语言 L_2 的可疑文本片段 s （$L_2 \neq L_1$）以及一种相关性排序方法 M，目标是从集合 D 中检索与 s 最相关的 K 段文本，其中 $K \ll |D|$。

该方法采用倒排索引实现快速搜索。参考文档被划分为更小的文本片段，然后进行预处理并建立索引。借助预定义的多语言同义词库（它将每个单词映射到一组唯一的概念），每个段落中的单词都被替换为对应的概念。未关联任何概念的单词保持不变。随后，为每个经过修改的段落建立倒排索引。可疑文档被划分为句子，并在句子层面进行替换。第 3.1.1 节将对此过程进行说明。

系统会针对替换后的可疑句子进行搜索，尝试找出与之最相似的源文档修改后段落。除了使用多语言同义词库外，为了提高搜索质量，我们还对源文档和可疑文档进行了预处理。文本经过词形还原后，所有停用词、数字、标点符号，以及其他内部含有标点符号的词元都会被删除。

上述文档索引和搜索流程使用开源企业搜索平台 Apache Solr[3] 实现。Okapi BM25 被用作排序函数 [30]。分词和词形还原使用了 Stanza v1.2.1 [31]。

候选检索过程如图 1 所示。

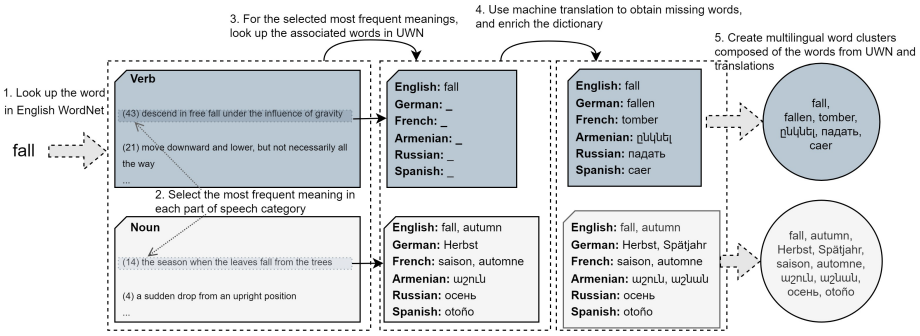

图 2：利用 Universal WordNet 创建初步多语言词簇，并通过机器翻译对其进行扩充的过程。

3.1.1. 多语言词簇

为了为每个词确定正确的概念，先前发表的研究依赖于词义消歧。本文提出绕过这一环节，采用基于词频的近似方法，将不同语言的词库合并为一个词库。

有许多多语言词库可用于上述方法。BabelNet [19], DBNary [32], 以及各种多语言 WordNet词网都开放了数据访问权限，这些数据可以加以改造，用于创建跨语言检索任务所需的多语言词簇。综合考虑词汇量、语言覆盖范围和精确率，我们决定以 Universal WordNet（UWN） [33] 作为实验基础。

在一种语言中，每个词都关联一个或多个同义词集，每个同义词集都对应一个独特的概念。将 UWN 中不同语言相互关联的同义词集合并后，所创建的多语言概念聚类所包含的词语数量会进一步增加。使用某个词的所有可用同义词集，会导致该词出现在许多聚类中，从而降低检索方法的精确率。因此，重要的是将每个词所属的聚类数量减少到最相关的几个。

[3] `https://solr.apache.org/`

在这项工作中，我们尝试了几种缓解这一问题的方法。首先，我们使用 Princeton WordNet 3.1[4] 中的概念来创建初始词簇集合，并为每个英语单词提取概念。随后，我们使用 UWN 将其他语言中对应的词语添加到这些词簇中。

设 L_{UWN} 为 UWN 支持的所有语言的集合， C_{UWN} 为所有语言中普遍存在的概念的集合， w 为某个英语单词。我们用 $S^l_{C_w}$ 表示语言 $l \in L$ 中与单词 w（$C_w \in C_{UWN}$）的所有概念相对应的同义词集集合。然后，我们可以将不同的多语言词簇创建方法定义如下:

全部。对于每个英语单词 w，我们合并所有 $l \in L_{UWN}$ 对应的同义词集 $S^l_{C_w}$。该单词 w 不同词性的同义词集分别合并。合并后的同义词集即为多语言词簇。

前1。在这种方法中，我们还会根据普林斯顿 WordNet 提供的频率筛选概念集合。对于每个英语单词 w，我们只选取频率最高的概念 C^{Top1}_w，并合并其对应的、所有 $l \in L_{UWN}$ 的同义词集 $S^l_{C^{Top1}_w}$。对于单词 w 的每个词性类别，分别选取频率最高的概念。

我们将创建的多语言词语到词簇映射器称为“跨语言同义词词典”（CL-SynDi）。由于 Universal WordNet 收录了 200 多种语言中的 1,500,000 多个词，CL-SynDi 映射器可用于构建适用于所有这些语言的跨语言信息检索系统。

对 UWN 得出的词到概念的映射进行分析后发现，许多概念在某些语言中缺少对应的词。因此，我们以翻译的方式，手动为相应语言中的词簇补充缺失的含义。此外，我们还用已收录语言中的词汇丰富了这些词簇。翻译使用 Google Chrome 翻译功能完成，仅以英语单词作为翻译源。为了获得那些可能

[4] https://wordnet.princeton.edu/

作为不同词性的词语的恰当译文，我们结合其词性进行翻译。CL-SynDi 的创建过程如图 2 所示。所得词典的最终统计数据见图 3。

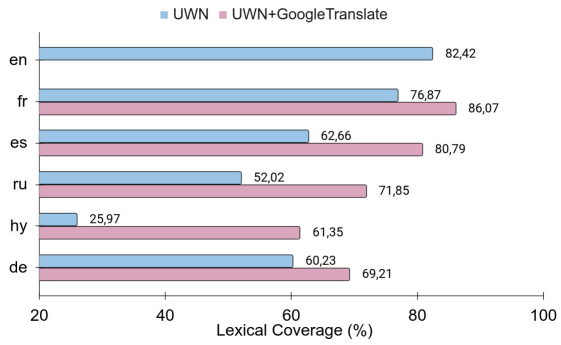

图 3：使用和未使用机器翻译创建的多语言聚类在词汇覆盖率方面的比较，基于维基百科的 120,000 篇文本计算得出。

虽然粒度不够细，但所提出的方法在各种跨语言基准测试中取得了良好结果；正如亚美尼亚语的案例所示，对于可能没有词义消歧工具的资源匮乏语言，该方法也相当有效。

创建的多语言聚类已发布在 GitHub[5] 上。

3.2. 详细分析

检索到前 K 个相关片段后，我们会进一步详细分析它们与可疑片段的相似度。此步骤旨在提高抄袭检测的整体精确率。

在详细分析步骤中，我们采用与 [34]相同的方法。该任务被定义为二元分类：判断两段不同语言的输入文本是否互为翻译。与 [34], 类似，我们也对预训练多语言掩码语言模型进行微调，并以一对句子作为输入。如果多个候选项都被分类为翻译，则选择分数最高的候选项。模型及其工作原理如

[5] https://github.com/1998karen/Cross-Lingual-Plagiarism-Detection

图 4. 如果检索到的文本片段和可疑文本片段包含多个句子，则将其拆分为句子，再使用微调后的语言模型逐对进行比较。

值得注意的是，所定义的翻译检测任务与释义检测相似。考虑到这一点，我们决定在任务中使用多语言 BERT 语言模型，因为它们在释义检测方面已取得不错的效果。

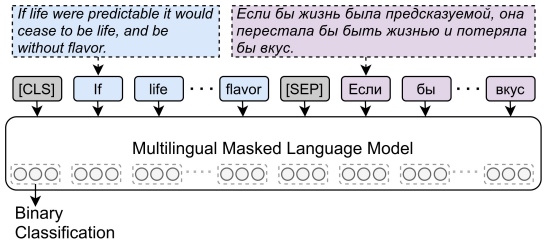

图 4：详细分析阶段用于判断两个文本片段是否互为翻译的分类过程。

为了微调语言模型，我们将平行句对用作正例，并将相似但彼此并非翻译关系的句子用作负例。

4. 实验

为展示所提方法的有效性，我们开展了多项实验，分别评估候选检索和详细分析阶段，并评估该方法在既有基准和新测试数据集上的整体表现。我们针对英语-俄语、英语-法语、英语-西班牙语、英语-德语和英语-亚美尼亚语语言对进行了评估。

4.1. 候选检索

为了从候选检索算法部分提出的改进方案和同义词词典中选出最佳方案，我们进行了多项测试。我们向索引中添加了 10,000

篇从维基百科提取的英文文档，作为参考语料库。为了评估候选检索的性能，我们随后检索了 316 篇自动生成的俄语可疑文档。这些可疑文档都包含抄袭片段。

候选检索步骤的评估指标选用$Recall@K$，该指标计算检索到的 K 文档中相关文档所占的比例。

根据表1所示的估算结果，使用仅基于频率得分最高的词义创建的词到概念映射器，取得了最佳结果。

	Recall@K			
合并方法 K=1		**K=5**	**K=10**	**K=50**
All	0,624	0,728	0,762	0,832
Top1	**0,747**	**0,827**	**0,853**	**0,899**

表1：候选项检索任务中跨语言叙词表合并方法的外部评估。

此外，我们还检验了通过机器翻译扩展叙词表的效果（表2）。

	Recall@K			
叙词表	**K=1**	**K=5**	**K=10**	**K=50**
UWN	0,175	0,265	0,307	0,426
UWN+Google 翻译	**0,747**	**0,827**	**0,853**	**0,899**

表2：机器翻译的多语言叙词表在检索任务中的有效性外部评估。

4.1.1. 超参数

在候选检索中，我们选择了 $K = 50$，以确保较高的召回率。例如，[35] 在英匈抄袭检测中，也在检索阶段考虑了最多 50 篇文档。对于候选文档的 Okapi BM25 排序，我们使用了 $k1 = 1.2$ 和 $b = 0.75$。可疑文档被划分为

句子，而参考文档则被划分为段落。此配置用于之后的所有实验。

4.2. 详细分析

我们选择 XLM-RoBERTa (XLM-R) [36] 作为模型来完成详细分析任务。XLM-RoBERTa 是一个使用 100 种不同语言训练的多语言模型。

我们创建了一个训练数据集，用于微调 XLM-R 模型，随后在多个数据集上评估其性能。

训练。为了微调 XLM-R，我们使用英文维基百科以及从 Google Scholar 随机抽取的各主题科学论文，自动生成了一个数据集。正例通过自动翻译科学论文中的随机句子和维基百科中的句子生成，两者各 4,500 个。训练集共包含 9,000 个正例，每个语言对 150 个。

对于负例，我们从英文维基百科和科学论文中为每个语言对分别随机抽取 75 对句子。随后将这些句子翻译成相应的语言，为每个语言对构成句子对。我们假设，在抄袭检测过程中，大多数进入详细分析步骤的句子对，其句子之间不会存在相似性，或相似度较低。在生成数据的过程中，我们检查并确保训练集中没有重复出现的句子。最终，训练集共包含 18,000 个样本。

评估。训练好的模型在现有的 Negative-1、Negative-4 测试数据集 [34] 和一个基于 MRPC 的新数据集 [37]上进行了评估。评估结果见**表3**。

Negative-1、Negative-4。这些来自 [34] 的测试数据集由英语-俄语句子对组成。它们仅用于针对该语言对微调模型。正例使用了来自 Yandex 平行语料库的 16,000 个句子和 4,000 个手动翻译的句子。负例则使用相同的 20,000 个句子生成。每个俄语句子都与

每个英语句子进行比较（其对应句子除外），并使用不同的相似度度量。对于 Negative-1 集，每个俄语句子仅从最相似的句子中选取一个负例。对于 Negative-4 集，则选取 4 个负例（每种相似度度量选取一个最相似的例子）。

νMRPC。现有细粒度分析数据集的主要缺点是，它们生成非翻译句对时主要依据词汇距离。这种方法遗漏了一些棘手的情况：句子在词汇上相距甚远，但语义却非常接近。鉴于释义检测任务与翻译检测任务之间的相似性，我们想到将释义识别数据集转换为适用于我们需求的数据集。对释义识别数据集中的近似释义进行翻译，可以弥补缺少具有挑战性的负例这一问题。为了更稳健地评估训练后的细粒度分析模型，我们基于微软研究院释义语料库（MRPC）的机器翻译，为 5 种语言对创建了一个新的测试数据集。我们将这个新数据集记为νMRPC。

图 5 描述了创建 νMRPC 的过程。对于 MRPC 数据集中的每个语言对，我们将其中一个句子翻译成 5种选定语言。翻译使用 Google 翻译完成。因此，每个未翻译的英语句子都与其对应原句的译文配对。此外，为排除低质量翻译对模型性能造成影响的风险，我们使用 ModelFront [6] 对所有源句与译句对进行了翻译风险预测。估计风险评分超过 50% 的译文均被舍弃。在最终得到的数据集中，各语言的平均风险评分介于 24% 至 35% 之间。

这种方法使我们能够生成更具挑战性的负例来评估模型。与 SemEval-17 Task 1 [38]，需要专家手动参与不同，我们能够自动为多种语言创建样例，其中包括亚美尼亚语这类资源匮乏语言，

[6] https://modelfront.com

数据集	语言	语言对	召回率	精确率	F1
νMRPC	英语-亚美尼亚语	901	0,89	0,74	0,81
	英语-俄语	1250	0,91	0,72	0,81
	英语-西班牙语	1457	0,94	0,72	0,82
	英语-法语	1397	0,91	0,72	0,81
	英语-德语	1154	0.94	0.70	0,81
负-1	英语-俄语	7998	0,89	0,93	0,91
负-4	英语-俄语	18613	0,84	0,88	0,86

表3：不同基准测试上详细分析模型的分数。

然后比较模型在不同语言上的检测质量。

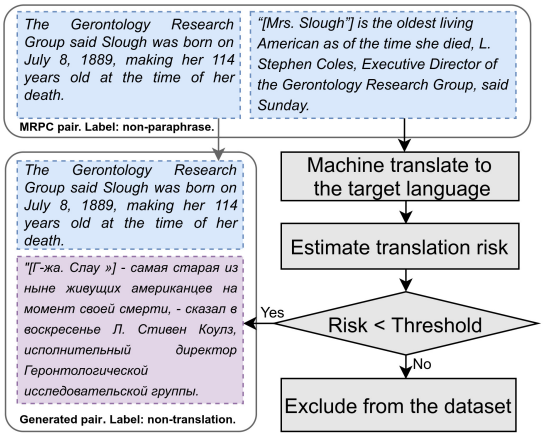

图 5：详细分析具有挑战性的测试数据生成，即翻译 MRPC 释义对中的其中一个句子。

在接下来的实验中，为了获得更高的精确率，微调模型的分类阈值根据最高的 $F_{0.25}$ 分数进行选择。

4.3. 跨语言抄袭检测基准

对于英语-俄语、英语-法语和英语-西班牙语语言对，我们使用了现有的跨语言抄袭检测基准，并将结果与这些基准上的新方法和最先进方法的结果进行了比较。我们还为所有这些语言对构建了一个新的测试数据集，并在其上评估了该方法。由于

	Crosslang		
	Rec.	**精确率**	**F1**
所提方法	0,77	**0,86**	**0,81**
[9]	**0,79**	0,83	0,80

表4：在英语-俄语跨语言抄袭检测任务的 CrossLang 数据集上，所提方法与最先进方法的比较。

本研究的范围不包括对旨在提高粒度分数的后处理步骤进行研究，因此 [39]，我们仅根据精确率和召回率指标比较各方法。下面介绍所选数据集和评估结果：

Crosslang 是一个自动生成的英语-俄语跨语言抄袭检测数据集 [9]。该数据集基于从英语维基百科和俄语维基百科中随机选取的文章。在此，待检测文档与对应的源文档主题相似。每篇待检测文档包含 20% 至 80% 的抄袭内容，其来源分散在 1 至 10 篇文档中。

我们将所提方法的性能与该基准上的最先进方法进行了比较。CrossLang 数据集上的评估结果见表4。

[40]。我们还使用了 [40] 预先收集的不同平行语料库子集，这些子集常用于不同的跨语言自然语言处理任务：

- *JRC-Acquis*[7] ——该语料库包含《共同体法律汇编》中以 22 种语言编写的欧洲法律文本。
- 亚马逊产品评论 ——这是一个由 [41]收集的不同产品评论语料库。
- 会议论文（*TALN*） [42] - 一个包含从以两种语言发表的会议论文中提取的科学文本的语料库。

[7]ec.europa.eu/../jrc-acquis

我们采用与 [13]相同的评估方法，并将我们的方法的性能与其论文中公布的结果进行比较（表5）。

	JRC-Aquis	APR	TALN
所提方法	**71,80**$_{\pm 0,444}$	**96,67**$_{\pm 0,387}$	**89, 72**$_{\pm 0,474}$
[13]	**72,70**$_{\pm 1,446}$	78,91$_{\pm 1,005}$	80,89$_{\pm 0,944}$

表5：所提方法与最先进方法在 Ferrero等人的英语-法语跨语言抄袭检测数据集上的F1分数比较。

PAN-PC-2011。该语料库包含单语和跨语言抄袭案例 [43]。PAN-PC-2011包含自动生成和人工生成的抄袭案例文档。对于跨语言抄袭检测任务，该语料库包含西班牙语-英语和德语-英语文档对。其中一些文档在部分文本经过自动翻译后经过了人工校正。为了使用该数据集评估我们的算法，我们仅使用为外部抄袭检测任务创建且用于跨语言案例的文档。

我们使用了 PAN 2010 会议组织者提供的评估脚本[8] 。西班牙语-英语子集的结果如表6所示。我们将自己的方法与 [28] 提出的最新最先进方法进行了比较。

	PAN-PC-11 ES-EN	
	召回率	**精确率**
所提方法	**0,79**	**0,85**
[28]	0,75	0,79

表6：所提方法与PAN-PC-11数据集英语-西班牙语子集上的最先进方法的比较。

所有语言对的数据集。为了评估和比较

[8]github.com/pan-webis-de/../sepln09

所提出的算法在5种选定语言上的效率，我们基于维基百科文章创建了一个新数据集[9]，采用与[9]相同的算法。每种语言的数据集都包含400篇可疑文本，其中含有从由120,000篇英文文本构成的参考语料库中抄袭的段落。

详细统计数据和评估结果分别见表7和表8。

	抄袭比例		
Pair	**几乎没有 (0-0.2)**	**中等 (0.2-0.5)**	**Much (0.5-0.8)**
英语-亚美尼亚语	16.5%	65.25%	18.25%
英语-俄语	15.25%	62.25%	22.5%
英语-西班牙语	7.5%	59.0%	33.5%
英语-法语	12.0%	60.0%	28.0%
英语-德语	19.0%	63.0%	18.0%

表7：所提出的跨语言抄袭检测数据集中每篇文档的抄袭比例。

Pair	**召回率**	**精确率**	**F1**
英语-亚美尼亚语	0,72	0,73	0,73
英语-俄语	0,81	0,82	0,81
英语-西班牙语	0,90	0,86	0,88
英语-法语	0,88	0,81	0,84
英语-德语	0,71	0,64	0,67

表 8：所提出的跨语言抄袭检测基准上的评估结果。

[9] https://drive.google.com/drive/folders/1jnAehDCQM_u1P3wKpRMozpbpiu0xbP5E?usp=sharing

5. 讨论

评估结果表明，基于最常见义项对多语言同义词集进行分组这一简单方法，在跨语言抄袭检测中非常有效。基于改进后的 UWN 和 XLM-R 的方法，在既有测试数据集上的表现接近英语-法语和英语-俄语语言对的最先进方法。

针对亚美尼亚语 MRPC 的详细分析结果表明，即使对于资源匮乏语言，多语言模型在这项任务中也能发挥作用。我们的方法在整体检测任务中也取得了良好结果，说明它适用于资源匮乏语言。所提方法的优势在于不依赖机器翻译，也不需要进行细粒度的词义消歧。因此，它可以轻松适配并应用于 UWN 和 XLM-R 支持的所有语言。

谈及这项工作的局限性，我们需要指出几点。首先，所提方法仅适用于具备词元化和分词工具的语言。其次，Universal WordNet 收录的语言数量有限，这也限制了可适用语言的数量。此外，由于该方法使用机器翻译工具来补充“跨语言同义词词典”，因此需要质量较好的机器翻译工具。再者，结合德语文本上的实验结果，我们可以推测，该方法在常用复合构词的语言中存在一些问题。值得注意的是，该方法仅在印欧语系的语言上进行了测试。

6. 结论

在这项工作中，我们提出了一种跨语言抄袭检测方法，该方法依托公开可用的多语言词库和预训练语言模型，因此可以轻松适配多种语言。该方法的有效性

既通过现有基准测试在大语种上得到了验证，也借助新提出的数据集在一种资源匮乏语言上得到了验证。

7. 致谢

本研究获得了俄罗斯联邦政府分析中心依据补助协议（协议编号000000D730321P5Q0002）及其与俄罗斯科学院伊万尼科夫系统编程研究所于2021年11月2日签订的第 70-2021-00142 号协议提供的人工智能领域研究中心资助。

作者感谢 Denis Turdakov、Yaroslav Nedumov 和 Kirill Skorn yakov 提供的宝贵意见。

参考文献

[1] M. Potthast、B. Stein、A. Eiselt、A. B.-C. P. Rosso，《第一届国际抄袭检测竞赛综述》，载于：第三届PAN研讨会。《揭示抄袭、作者身份与社交软件滥用》，2009年，第1页。

[2] M. Potthast, A. Barr´on-Cede˜no, A. Eiselt, B. Stein, P. Rosso，《第二届国际抄袭检测竞赛》概述，载于：M. Braschler、D. Harman、E. Pianta（编），《CLEF 2010评测实验室论文集》，《计算机科学讲义》第1176卷，2010年。

[3] M. Potthast, A. Eiselt, L. A. Barr´on Cede˜no, B. Stein, P. Rosso，《第三届国际抄袭检测竞赛》概述，载于：CEUR研讨会论文集，第1177卷，CEUR研讨会论文集，2011年。

[4] S. Muhammad，《单语与跨语言释义文本复用及外部抄袭检测》，博士论文，兰卡斯特大学（2020年）。

[5] C. K. Kent、N. Salim，基于 Web 的跨语言抄袭检测，载于：2010年第二届计算智能、建模与仿真国际会议，IEEE，2010年，第 199–204 页。[6] M. A. Sanchez-Perez、G. Sidorov、A. F. Gelbukh，一种在 PAN 2014 中用于文本复用检测的文本对齐获奖方法，载于：CLEF（工作论文），2014年，第 1004–1011 页。[7] I. Muneer、M. Sharjeel、M. Iqbal、R. M. A. Nawab、P. Rayson，Cleu——用于文本复用实验的跨语言英语-乌尔都语语料库与基准，《信息科学与技术协会学报》70（7）（2019年）729–741。[8] O. Bakhteev、A. Ogaltsov、A. Khazov、K. Safin、R. Kuznetsova，Cross-lang：跨语言抄袭检测系统，载于：NeurIPS 2019 文档智能研讨会，2019年。[9] M. Kuznetsova、O. Bakhteev、Y. Chekhovich，大型文本集合中的跨语言文本复用检测方法，Informatika i Ee Primeneniya[《信息学及其应用》] 15 （1）（2021年）30–41。[10] M. Potthast、B. Stein、M. Anderka，基于维基百科的多语言检索模型，载于：C. Macdonald、I. Ounis、V. Plachouras、I. Ruthven、R. W. White（编），《信息检索进展》，施普林格·柏林·海德堡，柏林，海德堡，2008年，第 522–530 页。doi:https://doi.org/10. 1007/978-3-540-78646-7_51。[11] A. Barr´on-Cede˜no、P. Rosso、D. Pinto、A. Juan，使用统计模型进行跨语言抄袭分析，PAN 212。[12] M. Franco-Salvador、P. Rosso、M. Montes-y-G´omez，知识图谱分析在跨语言抄袭检测中的系统研究，《信息处理与管理》52（4）（2016年）550–570。[13] J. Ferrero、F. Agnes、L. Besacier、D. Schwab，使用词嵌入进行跨语言抄袭检测，arXiv 预印本 arXiv:1702.03082。

[14]J. Devlin、M.-W. Chang、K. Lee、K. Toutanova，《BERT：用于语言理解的深度双向 Transformer 预训练》，arXiv 预印本 arXiv:1810.04805。[15] Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer、V. Stoyanov，《RoBERTa：一种鲁棒优化的 BERT 预训练方法》，arXiv 预印本 arXiv:1907.11692。[16] A. Conneau、G. Lample、M. Ranzato、L. Denoyer、H. J´egou，《无需平行数据的词语翻译》，arXiv 预印本 arXiv:1710.04087。[17] M. Potthast、A. Barr´on-Cede˜no、B. Stein、P. Rosso，《跨语言抄袭检测》，《知识型系统》45（2011年）45–62。[18] P. McNamee、J. Mayfield，《面向欧洲语言文本检索的字符 n 元语法分词》，《信息检索》7（2004年）73–97。[19] R. Navigli、S. P. Ponzetto，《BabelNet：覆盖范围广泛的多语言语义网络的自动构建、评估与应用》，《人工智能》193（2012年）217–250。[20] P. Vossen，《EuroWordNet 简介》，载于：《EuroWordNet：包含词汇语义网络的多语言数据库》，施普林格，1998年，第1–17页。[21]Z. Ceska、M. Toman、K. Jezek，《多语言抄袭检测》，载于：AIMSA，2008年。[22]P. Gupta、A. Barr´on-Cede˜no、P. Rosso，《基于概念词库的跨语言高相似度搜索》，载于：T. Catarci、P. Forner、D. Hiemstra、A. Pe˜nas、G. Santucci（编），《信息获取评测：多语言性、多模态性与可视化分析》，施普林格·柏林·海德堡，柏林，海德堡，2012年，第67–75页。[23] M. Franco-Salvador、P. Gupta、P. Rosso，《基于多语言语义网络的跨语言抄袭检测》，载于：P. Serdyukov、P. Braslavski,

S. O. Kuznetsov、J. Kamps、S. R¨uger、E. Agichtein、I. Segalovich、E. Yil- maz（编），《信息检索进展》，施普林格·柏林·海德堡，柏林，海德堡，2013年，第710–713页。

[24] M. Franco-Salvador、P. Rosso、M. Montes-y-G´omez，《基于知识图分析的跨语言抄袭检测系统研究》，《信息处理与管理》52（4）（2016年）550–570。

[25]E. Gabrilovich、S. Markovitch，使用基于维基百科的显式语义分析计算语义相关性，第6卷，2007年。

[26] P. F. Brown、S. A. Della Pietra、V. J. Della Pietra、R. L. Mercer，《统计机器翻译的数学：参数估计》，《计算语言学》19（2）（1993年）263–311。

[27] A. B´erard、C. Servan、O. Pietquin、L. Besacier，Multivec：用于自然语言处理的多语言多层次表示学习工具包，载于：第十届语言资源与评测会议（LREC），2016年。

[28] M. Roostaee, S. M. Fakhrahmad, M. H. Sadreddini，跨语言文本对齐：一种用于抄袭检测的双层匹配方案，《专家系统与应用》160 (2020年) 113718。[29] M. Roostaee, M. H. Sadreddini, S. M. Fakhrahmad，一种用于跨语言抄袭检测的有效候选检索方法：基于概念和关键词的方案融合，《信息处理与管理》57 (2) (2020年) 102150。

[30] S.E. Robertson, S.Walker, S.Jones, M. M. Hancock-Beaulieu, M. Gatford，Okapi 在 TREC-3 中的应用，美国国家标准与技术研究院特别出版物 SP 109 (1995年) 109。

[31] P. Qi、Y. Zhang、Y. Zhang、J. Bolton、C. D. Manning，《Stanza：面向多种人类语言的 Python 自然语言处理工具包》，载于：第58届计算语言学协会年会论文集：系统演示，2020年。

[32] G. S´erasset, 《Dbnary：基于 lemon 的 Wiktionary 多语言词汇资源》，RDF，《语义网》6（4）（2015）355–361。[33] G. de Melo、G. Weikum，《通过综合证据学习构建通用 WordNet》，载于：D. W. Cheung、I. Song、W. W. Chu、X. Hu、J. J. Lin（编），《第18届ACM信息与知识管理会议论文集》（CIKM 2009），ACM，美国纽约州纽约市，2009年，第513–522页。doi:http://doi.acm.org/10.1145/1645953.1646020。[34] D. Zubarev、I. Sochenkov，《基于上下文模型和上下文无关模型的抄袭检测跨语言文本对齐》，载于：年度国际会议“对话”论文集，第1卷，2019年，第799–810页。[35] M. Pataki，《搜索翻译抄袭的新方法》。[36] C. Alexis、K. Kartikay、G. Naman、C. Vishrav、W. Guillaume、G. Francisco、G. Edouard、O. Myle、Z. Luke、S. Veselin，《大规模无监督跨语言表示学习》（2020年）。arXiv:1911.02116。[37] W. B. Dolan、C. Brockett，《自动构建句子释义语料库》，载于：第三届国际释义研讨会（IWP2005）论文集，2005年。[38] D. Cer、M. Diab、E. Agirre、I. Lopez-Gazpio、L. Specia，《SemEval-2017任务1：语义文本相似度——多语言与跨语言专项评测》，arXiv预印本 arXiv:1708.00055。[39] M. Potthast、B. Stein、A. Barr´on-Cede˜no、P. Rosso，《抄袭检测评估框架》，第2卷，2010年，第997–1005页。[40] J. Ferrero、F. Agn`es、L. Besacier、D. Schwab，《用于跨语言文本相似度检测的多语言、多风格、多粒度数据集》，2016年。[41] P. Prettenhofer、B. Stein，《使用结构对应学习进行跨语言文本分类》，载于：第48届计算语言学协会年会论文集，2010年，第1118–1127页。

[42] F. Boudin, 《TALN档案库：自然语言处理领域法语研究论文的数字档案库》（TALN档案库：法语研究论文的数字档案库）[法语], 载于：《TALN 2013会议论文集》（第2卷：短篇论文），2013年，第507–514页。

[43] M. Potthast、B. Stein、A. Eiselt、A. Barr´on-Cede˜no、P. Rosso, 《PAN抄袭语料库2011》（PAN-PC-11）（2011年6月）。doi: https://doi.org/10.5281/zenodo.3250095。