arXiv:2601.00488v1 [cs.CL] 1 Jan 2026

面向历史VET文档的噪声感知命名实体识别*

Alexander M. Esser[1] 和 Jens D¨orpinghaus [1,2]

[1]德国波恩联邦职业教育与培训研究所（BIBB）[2]德国科布伦茨大学计算机科学系 alexander.esser@bibb.de, doerpinghaus@uni-koblenz.de

摘要

本文探讨职业教育与培训（VET）领域的命名实体识别（NER），重点关注受到OCR引入的噪声影响的历史数字化文档。我们提出一种鲁棒的NER方法，结合噪声感知训练（NAT）、合成注入的OCR错误、迁移学习和多阶段微调。我们系统比较了三种互补策略：使用含噪数据、干净数据和人工数据进行训练。我们的方法是首批能够识别职业教育与培训文件中多种实体类型的方法之一。该方法应用于德语文档，但也可迁移至任意语言。实验结果表明，领域特定的噪声感知微调显著提升了含噪条件下的鲁棒性和准确率。我们提供可公开获取的代码，以便在特定领域背景下复现噪声感知命名实体识别。

关键词：命名实体识别（NER）、噪声感知训练（NAT）、数据增强、职业教育与培训（VET）、OCR噪声、历史文档、文档处理、信息抽取、BERT。

1 引言

本文提出了一种稳健的方法，用于劳动力市场研究领域的命名实体识别（NER），尤其关注职业教育与培训（VET）。该领域既包含当代数据，也包含历史数据；本文将聚焦于历史数字化文档。我们采用噪声感知训练（NAT），以应对光学字符识别（OCR）所引入的噪声这一挑战。

*这是在 VISAPP 2026 上发表的论文的扩展版，未经同行评审。

1.1 研究动机

劳动力市场不断变化，受到全球发展趋势和各国特定条件的影响。在德语国家，数字化等技术创新以及社会变迁要求员工掌握新的技能。具体而言，职业教育与培训（VET）、再培训以及继续职业教育与培训（CVET）是满足这些需求的关键（参见 Dobischat 等，2019；Helmrich 等，2016）。德国的政策制定者日益认识到，继续职业培训和终身学习对于数字化转型的成功至关重要（参见 Schiersmann，2022年）。

该体系将职业教育与培训作为专业人士高等教育的一部分提供（*Aus-bildung*、*Umschulung*）。此处的职业教育与培训一词也涵盖继续职业教育与培训，包括进阶培训和提升培训（*Weiterbildung*、*Fortbildung*）。

联邦职业教育与培训研究所（BIBB）拥有与职业教育与培训相关的大型数据集，其中包括数字化档案文献。Reiser 等人（2024）发布了一个历史VET文档数据集，涵盖从1908年到现在的文档，共计2125份。

分类系统与分类体系

国际职业分类标准（ISCO）由国际劳工组织（ILO）制定，并于1958[1]首次发布。此后分别于1968年和1988年推出了新版，最新版本于2008年发布。该标准也已被欧盟（EU）采用，并针对一些德语国家制定了特定版本，包括德国、奥地利和瑞士。ISCO中的职业按技能等级划分，并与欧洲技能、能力、资格和职业分类（ESCO）本体相连，为数据增添了另一个层级（参见 Reiser 等人，2024；Dorau 和 Hein，2025）。

在德国，联邦就业局（BA）及其研究机构 IAB 所采用的参考分类是德国职业分类系统 （职业分类；KldB）。该系统按照任务层级对职业进行组织。最新版本是 KldB 2010 的2020年修订版，该版本经过全面重新设计，因此1988年和1992年发布的旧版本已不再适用。该系统旨在与 ISCO-08 兼容。德国劳动力市场本体 （GLMO）收录了德国使用的所有分类（参见 Fischer 和 D¨orpinghaus，2024；D¨orpinghaus 等，2023）。

职业教育与培训研究议题

对多种劳动力市场数据源进行自动化处理和整合，被认为是一项重要但复杂的工作（参见 Fischer 和 D¨orping-haus，2024）。我们的长期目标是处理并构建职业教育与培训数据集，使其能够用于数据仓库或构建为知识图谱。这样，研究人员便能探索职业教育与培训及更广泛的社会科学领域中的各类相关问题，也能研究心理学和经济学等领域的问题。

基于本研究识别出的实体，可以建立职业教育与培训文件之间的引用关系，并将其表示在知识图谱中。在此基础上

[1]参见 https://www.ilo.org/public/english/bureau/stat/isco/isco08/。

，可以进一步探讨职业教育与培训领域的研究问题，例如比较西德和东德的历史培训法规、职业资格的认证，或职业画像随时间的演变。

文档处理流水线

为实现这一目标，需要一个文档处理流水线来自动处理大量《数据》并进行信息抽取。遵循经典架构（参见 Konya，2013），此类处理流水线通常由各种预处理步骤、中央光学字符识别引擎以及后续的信息抽取（IE）任务组成，例如命名实体识别（NER）。

一个主要挑战在于，OCR错误和领域专用术语会显著影响命名实体识别（NER）等下游任务。因此，我们提出了一种 NAT 方法，以提高命名实体识别（NER）对 OCR错误的鲁棒性。

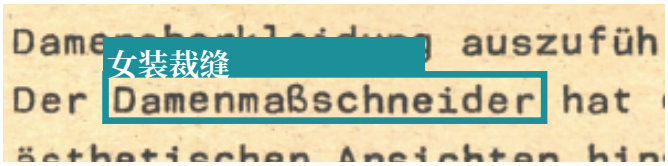

图 1：职位名称中的 OCR错误示例。

图 1 展示了本工作的核心目标示例：在历史职业教育与培训文档的页面上，应识别出各种实体。术语“女装裁缝”（女装裁缝）在光学字符识别过程中被错误拼写；然而，在命名实体识别（NER）过程中，仍应将其识别为一个职位名称 实体。

1.2 贡献

本文的贡献如下：

- 我们提出了一种适用于职业教育与培训文件的稳健、噪声感知命名实体识别方法。
- 我们比较了领域特定微调与噪声感知训练策略，通过训练三种模型：含噪、干净和人工。
- 我们采用迁移学习进行领域适应，并通过多阶段微调将分类体系作为额外训练数据整合进来。
- 我们提出了首批能够识别《数据》中多种实体类型的方法之一，而以往的研究通常只关注单一实体类型。
- 随后，我们对不同实体类型的识别率、混淆情况及内在难度进行了定性分析。
- 我们对历史职业教育与培训文件中典型的OCR错误进行了定性误差分析。
- 为确保科学结果的可复现性，我们公开提供用于训练和评估的代码[2]。

[2]代码可在 https://github.com/TM4VETR/noise_aware_ner_vet/ 获取。

1.3 概述

本文其余部分的结构如下：第2节介绍文档处理流水线以及需要识别的实体的背景信息。第3节讨论命名实体识别、NAT 和职业教育与培训领域的相关工作。第4节介绍我们的方法，包括训练数据、错误注入和多阶段微调的详细信息。第5节评估三种互补的模型变体含噪、干净和人工，以及各实体类型的准确率。第6节总结全文并展望未来工作。

2 背景

本节概述了文档处理流水线，以及本研究所考虑的实体类型。

2.1 文档处理流水线

这项工作是我们更广泛研究计划的一部分。我们的长期目标是为职业教育与培训文件构建文档处理流水线。该流水线的首个版本包含光学字符识别和命名实体识别模块。图 2 展示了该处理流水线。

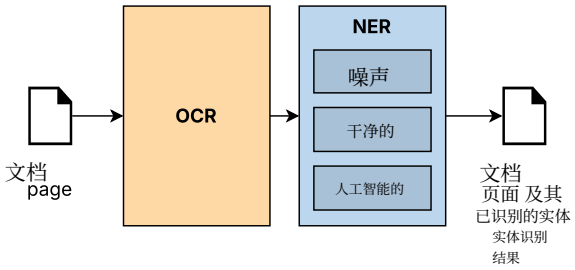

图 2：文档处理流水线。

文档通常分为两类：原生数字 文档（大多为 PDF 格式）和扫描文档（为 PDF 格式或各种图像格式）。许多方法只能处理原生数字 PDF 文档；在这种情况下，不会出现质量问题或 OCR错误。然而，要处理历史职业教育与培训数据，也必须处理扫描文档，因为这类文档占数据的绝大部分。

对于给定的文档页面（即扫描页面图像），首先使用*Tesseract* 光学字符识别引擎进行光学字符识别（Smith，2007）。随后进行命名实体识别。在此步骤中，会使用三种不同的模型：含噪、干净和人工模型，具体内容将在第4.1节中详细介绍。由此可为每个文档页面生成一份识别出的实体列表。

2.2 命名实体

进行实体识别时，需要在德语文本中识别以下命名实体：

1. 职位名称（JOB_TITLE）2. 职业名称组（JOB TITLE GROUP）3. 技能（SKILL）4. 培训科目（SUBJECT）5. 工作活动（ACTIVITY）

职位名称（JOB_TITLE）

作为实体 JOB_TITLE，需要识别具体的职业名称。在已标注的数据中，最常见的职位名称是“熟练工人”（技术工人）。

本文的实体识别任务并非根据分类体系对给定的职位名称进行分类，而是在自由文本中识别所有 JOB TITLE 实体。 -

职业名称组（JOB TITLE GROUP）

实体“职位类别”指包含多个相关职位名称的较高层级职业类别。该实体旨在将职业的层级结构纳入考量。

德国职业分类系统 KldB 定义了由五位数字代码表示的五级类别层级结构。第一位数字表示总体职业领域（职业领域），第五位数字则具体规定职业活动的要求等级或复杂度（职业类别）。

每个具体职位名称都被分配到这五个类别中最低层级的类别，例如，药学技术助理 这一职业（药学技术助理；代码 81822-105）被归入 81822 类别（参见 Bundesagentur f¨ur Arbeit，2020）：

8：卫生、社会服务、教学和教育
81：医疗卫生职业
818：药学
8182：药学技术助理职业
81822：药学技术助理职业——专业活动

标注数据中最常见的职位类别是“Facharbeiter- berufe”（技术工人职业）。

技能（技能）

技能是完成某种职业所需的心理或身体前提。在德语中，潜在的、较为一般的技能（*F¨ahigkeiten*）与习得技能（技能）之间存在细微区别。根据能力的教育学理解，这两类技能均辅以

知识后构成能力（参见 Hartig，2008年；Linten 和 Pr¨ustel，2015年；Weinert，2001年）。实体 SKILL 将这两类技能合并在一起。

识别技能时的一个特殊挑战是存在多词表达式，这些表达式要么以非常宽泛的术语描述某项技能，要么代表多项技能的并列组合（参见 Nguyen 等人，2024）；例如，“Aufmerksamkeit, Geduld und Fingerspitzengef¨uhl”（专注力、耐心和圆通）。最常被提及的技能是“专注力”（*Aufmerksamkeit*）。

培训科目（学科）
培训科目是职业教育中的课程主题（*Ausbil-dungsf¨acher*、*Lehrg¨ange*），通常列于课程大纲（课时表）中。作为实体“学科”，应识别具体的课程名称，而非详细的课程内容。标注数据中最常见的学科是“Werkstoffe”（工作材料）。

工作活动（活动）
实体“活动”指特定职业的具体工作活动。只有具体的行动才应标注为“活动”，而非泛指性短语。最常被提及的工作活动是“Reklamationen bearbeiten”（处理投诉）。

3 相关工作

本节回顾命名实体识别、自然语言处理以及职业教育与培训应用领域的现有方法。

3.1 命名实体识别（命名实体识别）

命名实体识别旨在根据预定义类别识别文本中的命名实体（参见 Nadeau 和 Sekine，2007）。命名实体识别可视为序列标注问题的一种特殊情况，该问题通常涉及为输入序列中的词元进行标注（参见 Tjong Kim Sang 和 De Meulder，2003）。传统的基于启发式的方法 与基于深度学习的方法 命名实体识别方法有所区别。

基于启发式的方法
在基于启发式的方法中，还可进一步区分基于规则的方法 与统计 方法：基于规则的方法包括模式匹配，依靠手工制定的规则或正则表达式来识别实体（参见 Eftimov 等，2017）。这些方法需要明确的词典，即预先编制的已知实体列表。统计方法，例如隐马尔可夫模型（HMMs）（Bikel 等，1999）和条件随机场（CRFs）（McCallum 和 Li，2003），采用基于人工设计特征的序列标注技术（参见 Jurafsky 和 Martin，2025）。

隐马尔可夫模型（HMMs）通过转移矩阵对序列进行建模，转移矩阵表示隐藏输入状态与可观测输出状态之间转移的联合概率（参见 Rabiner，1989）。条件随机场（CRFs）是无向图模型，用于计算给定相应输入序列时标签序列的条件概率（参见 Lafferty 等，2001）。诸如 Viterbi 算法（Viterbi，1967）

通常用于确定 HMMs 和 CRFs 中最可能的状态或标签序列。

基于深度学习的方法

相比之下，deep learning-based method 采用基于深度神经网络（DNN）的现代学习技术。这些方法还可根据模型架构进一步区分：基于 *LSTM* 的模型，尤其是双向 LSTM，广泛用于捕获上下文，并且通常与条件随机场结合进行序列解码（参见 Huang 等，2015；Lample 等，2016）。这些方法通常需要词嵌入和字符嵌入，例如 GloVe、FastText 或 FLAIR（参见 Pennington 等，2014；Bojanowski 等，2017年；Grave 等，2018年；Akbik 等，2018年）。

基于 *Transformer* 的模型，例如 BERT 及其众多变体，目前代表了最先进的水平（参见 Devlin 等，2019）。Transformer 架构由堆叠的自注意力层组成，用于建模词元之间的成对依赖关系，使每个词元都能获取整个序列的信息，同时在训练期间保持高度并行化（参见 Vaswani 等，2017年）。这些模型不需要显式词典，而是依赖从大型语料库中习得的上下文理解。因此，通常会在领域特定数据上对其进行微调，使模型能够隐式地学习并识别相关实体。

3.2 噪声感知训练（NAT）

噪声感知训练（NAT）是一组旨在使下游自然语言处理（NLP）系统（如命名实体识别）能够抵御上游引入的噪声（例如 OCR错误）的方法。NAT 的核心思想是向训练数据中注入逼真的错误，使模型学到对这类扰动不那么敏感的表示（参见 Namysł 等，2020）。在实践中，这包括注入各种特定于 OCR 的错误，例如字符替换、删除、插入和间距错误。

Namysł 等（2020）提出了一种用于序列标注的 NAT 方法，通过使用人工扰动，提升模型在命名实体识别任务中抵御 OCR错误和拼写错误的鲁棒性，同时不牺牲模型在干净文本上的性能。

Xu 等（2021）在文本分类之前注入合成 OCR 噪声，使模型能够处理含噪转录文本。

NAT 与经验误差建模密切相关；后者通过经验分析典型的 OCR错误，以便随后注入逼真的扰动（参见 Namysł 等，2021）。

文档噪声建模 通过显式建模文档处理场景中常见的失真，补充了噪声感知技术。*TrOCR* （Li 等人，2021）在其编码器—解码器架构的预训练过程中引入合成的文档退化（如模糊、低分辨率和扫描伪影）。类似地，OCR-free 模型（如 Donut（Kim 等人，2022年））会对文档图像施加可控扰动，使模型能够在含噪条件下直接从像素中学习文本提取。

自监督 *OCR* 校正超越了 NAT，旨在明确地学习纠正含噪 OCR 输出。近期方法采用去噪

序列到序列架构，通常基于 BART 或 T5 等预训练语言模型（参见 Singh 等人，2019；Saluja 等人，2021；Chen 等人，2021）。其他方法则利用置信度感知目标（参见 Boubt 等人，2022）或掩码语言建模来优化含噪 OCR 文本（参见 Huang 等人，2022）。

3.3 在职业教育与培训中的应用

针对劳动力市场数据，尤其是职业教育与培训文件中的命名实体识别，相关出版物数量有限。大多数出版物关注职位名称或技能的识别。

职位名称

Reiser 等人（2025年）检测并分类了德语文本中的职业，并比较了基于规则的方法与基于语言模型的方法。

Safikhani 等人（2023年）对 BERT 和 GPT-3 进行了微调，以自动抽取德语职业信息。

Decorte 等人（2021）提出了 JobBERT，这是一种基于深度学习的模型，用于识别职位名称并将其映射到 ESCO 类别。该方法以预训练模型 BERT 为基础，依据技能信息进行识别，表明技能描述是职位名称识别的重要组成部分。

需要指出的是，本文的目标不同于职业编码。职业编码是一项文本分类任务，旨在将给定的职位名称与预定义列表进行匹配，例如 KldB（参见 Gweon 等人，2017年；Schierholz，2014；Schierholz 和 Schonlau，2021）。相比之下，本文旨在从自由文本中识别职位名称（及其他实体类型）。

技能

Wang 和 Sun（2025年）提出了一种从招聘广告中抽取技能的处理流水线。

Li 等人（2023年）和 Nguyen 等人（2024年）近期采用大语言模型进行技能抽取。

Tabares 等人（2018年）提出了构建基于能力的本体的方法。

Huang 等人（2020年）提出了*Skill2Vec*，这是一种将技能数值化嵌入的方法，突破了 ESCO 专属建模的限制。

技能识别与能力建模密切相关，后者在教育研究中有着悠久的传统。相关工作包括一些基础框架，例如 O*NET（2020）、HR-XML 能力模式（2007）和 IEEE 可复用能力定义标准（2008）。关于德国能力模型，Tiemann 等人（2024）、Weinert（2001）、Hartig（2008）以及 Linten 和 Pr¨ustel（2015）提供了概述。

大多数出版物使用命名实体识别来识别单一类型的实体，例如职位名称或技能。它们没有考虑不同实体类型之间的语义联系。相比之下，本研究有意识别多种实体类型。搜索职位名称等明确表达容易出错；我们的长期目标是纳入额外的上下文信息和隐含信息，以提升待构建处理流水线的信息抽取能力。

表 1：按模型划分的训练数据量。

模型	实体	词元
含噪	1 058	7 280
干净	1 058	7 280
人工	2 116	14 560

4 提出的方法

OCR噪声会显著影响命名实体识别等下游任务。因此，我们提出一种 NAT 方法，以增强命名实体识别对 OCR错误的鲁棒性。

Namysł（2023年）深入研究了文档处理流水线的鲁棒性。领域特定再训练，即在一组领域特定文档上微调现有模型，通常已经能带来显著提升。更高级的方法是*NAT*。

实体的多变性带来了另一项挑战。其中一个原因是实体会随时间演变。这些历史VET文档来自不同年代，也来自两个社会政治制度不同的德国各州。术语（如职位名称）和表述方式（如活动描述）可能会随时间发生变化。语言特性也会带来更多差异。Reiser 等人（2025年）以职业名称为例，探讨了德语的独特挑战。职位名称的性别特定形式是造成差异的一个典型来源。

鉴于这些挑战，基于规则的方法并不理想。相反，我们训练了三种基于 Transformer 的模型，以 BERT 这一预训练的最先进模型为基础，专门针对识别德国职业教育与培训文档中的实体进行了适配，并使其能够捕捉该领域的词汇和特性。

4.1 提出模型

所提出的方法涉及训练三个模型：含噪、干净和人工模型。含噪模型基于含有未纠正 OCR 错误的训练数据。对于干净 模型，会在训练前纠正 OCR 错误。人工 模型则以干净 模型为基础，应用 NAT 技术，合成插入逼真的 OCR 错误（插入、删除、替换）。

请注意，这会导致训练数据量有所不同。含噪 和干净模型使用的原始训练数据量相同。对于人工 模型，每个拼写正确的单词都会被注入一种典型错误——替换、删除或插入 。这使数据量翻倍。表 1 列出了三个模型各自使用的训练数据量。

为确保可比性，本实验仅适度增加了训练数据量。然而，在实际应用中，NAT 提供了一种简单的方法，可通过数据增强来扩大数据集规模。

迁移学习

在模型训练过程中，会应用迁移学习。迁移学习是一种

表2：已标注实体数量。

实体	**1509**
职位名称	916
职位类别	121
技能	54
学科	167
活动	251

常见概念，即将在大型数据集上训练过的预训练模型，在较小的特定领域数据集上针对目标任务进行微调。Namysł 和 Konya（2019）证明，只需添加少量数据，就能显著改进骨干模型（在他们的案例中用于光学字符识别）。

4.2 训练数据

训练数据由来自东德、西德和统一后的德国的历史VET文档组成，尤其是培训条例（培训条例 / 职业条例）和职位名称的系统分类清单（系统分类清单），涵盖1976年以来的时期，总计68个文档页，并标注了OCR错误和实体识别。表2列出了标注数据中每种实体的出现次数。

理想的训练数据应包含所有可能的职位名称、职位类别、技能、学科和活动，且每种最好出现多次，既包括存在OCR错误的情况，也包括没有OCR错误的情况，这需要数百万个观测值。由于这无法实现，目标是提供一个足够大且真实的数据集，使模型能够从这些数据中进行泛化，以便用于未来的预测。

额外训练数据

为提高准确率，已整合额外的训练数据：

对于职位名称，使用了联邦就业局提供的搜索词列表[3]，，该列表将自由文本职业名称关联到 KldB，作为额外数据。

对于职位名称组，使用了 KldB（联邦就业局，2020），其层级结构包含大量表示职业的上位术语。

对于技能，使用了 ESCO分类法[4]（欧盟委员会，2025年），其中包含 30000 多项标准化技能。

然而，需要注意的是，所有这些额外训练数据在两个重要方面都不同于现实世界数据：每个词都属于某个实体；数据中没有“O”标签来表明相应词元不属于任何实体。此外，数据中没有 OCR噪声。

因此，额外训练数据不能直接用于最终模型。相反，我们采用多阶段微调，在中间阶段引入一个

[3]参见 https://www.arbeitsagentur.de/institutionen/dkz-downloadportal。[4] 参见 https://esco.ec.europa.eu/en/use-esco/download；版本 1.2.0。

表3：最常见的 OCR错误。

识别正确类型			频率
	-	删除	311
	,	删除	240
	e	删除	106
	.	删除	104
	n	删除	77
	r	删除	60
	t	删除	49
	s	删除	48
	u	删除	47
	i	删除	40
n	m	替换	39
	a	删除	36
ä	a	替换	33
i	l	替换	32

预训练阶段，并使用这些数据进行训练，如第4.4节所述。类似地，Schierholz 和 Schonlau（2021）通过加入来自索引的数据，提高了算法的性能（他们的研究针对职业编码）。

4.3 错误注入

使用 NAT 时，会在每个拼写正确的单词中注入一种错误：替换、删除或插入。

注入这些人工错误时，会明确考虑 OCR 引擎常见的错误；此类错误被插入的概率更高。为此，我们进行了实证错误分析。

实证OCR错误分析

定性错误分析显示，在历史VET文档中，以下类型的 OCR 错误尤为常见：(i) 未能识别标点符号（最常见的错误）；(ii) 将单个字母识别错误（例如，将“l”识别为“i”）；(iii) 难以识别数字（例如，将“0”识别为“o”）；(iv) 扫描伪影，即多出的单个字符（例如，误入的逗号或撇号）；(v) 将句点与逗号混淆（可能是由打字字体和文档老化所致）。表3列出了在经过标注的历史职业教育与培训文档中观察到的15种最常见 OCR 错误。

4.4 多阶段微调

为了整合第4.2节所述的额外训练数据，我们引入一个中间预训练阶段，并应用领域自适应预训练（*DAPT*）（参见 Gururangan等人，2020）。

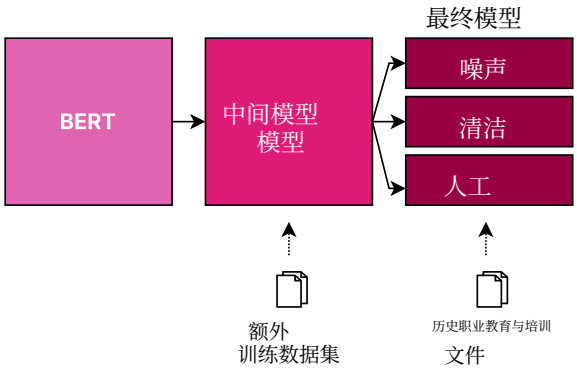

图 3：多阶段训练过程。

图 3 展示了多阶段训练过程：我们使用一个德语、区分大小写的 BERT 变体 dbmdz/bert-base-german-cased[5], 作为基础模型，该模型由巴伐利亚州立图书馆发布，其权重初始化自原始 BERT 的权重。

接着，使用大量额外训练数据对中间模型进行训练。这构成了使模型适应职业教育与培训领域的第一步，即应用领域自适应预训练，以专门识别第2.2节中提及的实体。最后，使用标注的真实世界数据训练三种模型变体：含噪、干净和人工。

Hendrycks等人（2019）表明，以这种方式对模型进行微调可以显著提升鲁棒性。由于OCR错误，这一点对于本用例尤为重要。

4.5 实现细节

该框架基于PyTorch（Paszke等人，2019）和Hugging Face Transformers库（Wolf等人，2022年）构建。

由于我们采用了预训练BERT模型，计算成本可以忽略不计；训练可在配备NVIDIA GPU的本地工作站上于短时间内完成。

标注数据集被划分为训练集、测试集和验证集（70:20:10），且各集合中的实体分布相同。划分在片段级别进行，而非词元级别，以避免拆分多词实体。

Bias

使用预训练模型时，需要注意这些模型可能会受到其训练数据中偏差的影响，从而影响预测结果。这种偏差也会延续到微调后的模型中。此外，额外的训练数据也可能存在偏差（参见 Wang和Russakovsky，2023年）。

[5]参见 https://huggingface.co/dbmdz/bert-base-german-cased。

超参数
训练最后三个模型时，我们使用以下超参数搜索验证损失的最低值：学习率调度与基于 Transformer 的模型尤其相关，因为这类模型通常对较大的参数变化很敏感。在预热阶段（占步骤的 10 %），学习率逐渐提升至目标速率 $2 \cdot 10^{-5}$。在随后的衰减阶段，随着训练的进行，学习率会逐渐降低至最小值（遵循逆根函数）。尤其是在微调预训练模型时，较小的学习率有助于进行更精细的调整。

最大训练轮次数为25。不过，分析（见第5.1节）表明，这一上限从未达到：损失函数在大约四个训练轮次后开始收敛（另见图4）；如果连续五个训练轮次均未观察到进展，则应用提前停止。

保存所有训练轮次中表现最佳的模型，评估指标为F_1 分数（在实体级别进行微平均）。

类别加权和过采样
为弥补显著的类别不平衡，我们采用类别加权，并额外使用过采样。大多数标签都是“O”，表示相应词元不属于任何实体。损失函数会被这些标签主导，模型可能退化为只预测“O”（类别坍缩）。对于类别加权，我们采用平衡（逆频率）加权，然后对权重进行归一化，并将其限制在一定范围内，以避免出现极端值。在训练过程中，过采样会复制包含正标签（非“O”）的词元窗口，从而提高提供给模型的正例相对频率。这种方法通过人工调整类别分布，增强了稀有实体的损失信号。

在对中间模型进行预训练时，采用了略有不同的超参数和设置：中间模型的额外训练数据不包含“O”标签。因此，没有采用过采样、权重平衡或截断。由于中间模型的数据集相当庞大（包含 557749 个词元—标签对），验证集被限制为 10 000 个词元。

标注：BIO格式
标注实体时，使用 *BIO* 格式，为每个词（词元）分配一个标签。标签“B”表示该词元标记实体跨度的起始位置；“I”表示该词元位于实体跨度内部；“O”表示该词元位于任何实体跨度之外。

词元简化
关于词元简化，存在两种相互矛盾的观点。一些作者建议在命名实体识别前对词元进行归一化（转换为小写、词干提取、移除特殊字符和停用词），以减少噪声（参见 Gweon 等人，2017年；Baldwin和Li，2015）。相反，另一些作者认为，过度预处理会损害命名实体识别，因为这会移除重要线索（参见 Mayhew等，2020；Alzahrani和Jololian，2021）。例如，大写形式可以为姓名和组织机构提供有用线索；形态学词尾（在德语中，例如 *-in*、*-meister*，以及复数形式中的 *-en*）有助于预测实体边界；词干提取则会抹平这些词尾。我们没有进行词元简化，因为实验表明，在我们的使用场景中，这样做并无益处（F_1 分数

从 77.5 % 降至 63.3 %）。

5 评估

本节将介绍评估结果，包括对三个最终模型的比较，以及它们在不同实体类型上的表现。

无法与 BERT 等最先进的命名实体识别模型直接比较，因为这些模型针对不同的实体类型（人物、地点等）进行训练，而非针对 VET 实体。

同样，也很难与其他使用 VET 数据的方法进行比较。第 3.3 节中提到的一些作者公布了职业名称识别的识别率；然而，他们使用的是不同的数据集，且任务是将职位名称与分类体系进行匹配，而不是在连续文本中识别职位名称。因此，这些识别率无法进行比较。基于此，我们报告性能指标仅供参考，而非用于建立基准。

命名实体识别的常见评估指标包括精确率、召回率和F_1分数。这些指标基于真正例 TP、假正例FP 和假负例 FN 的数量：

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$F_1 = \frac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$$

5.1 随时间变化的精确率和召回率

图4展示了含噪模型在训练过程中的精确率、召回率和F_1 分数。在整个训练轮次中，精确率和召回率始终接近，没有明显偏差。召回率持续略高于精确率。

从理论上讲，较高的召回率表明模型能够在句子中定位实体，但无法准确捕捉其边界。

5.2 模型比较

为确保可比性，三个模型均使用相同的参数，并在相同的训练集、测试集和验证集上进行训练。训练数据仅在其特征上有所不同：使用未经校正的光学字符识别结果（含噪 模型）、使用人工校正的光学字符识别结果（干净 模型），或注入合成类OCR错误（人工 模型）。

表4报告了各模型的准确率。人工 模型的F_1 分数最高（77.9%）。

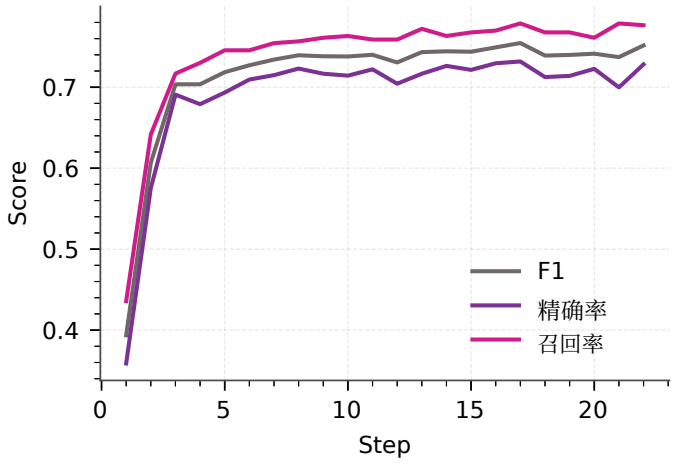

图 4：精确率、召回率和F_1 分数随时间的变化。

请注意，OCR校正和合成注入噪声只会影响训练集和验证数据集，不会影响测试数据集。之后在推理过程中，模型将接收含噪输入；因此，测试数据应能反映现实场景表现。

因此，正如预期，干净 模型的F_1 分数低于含噪模型，因为干净模型并未经过处理含噪数据的训练。这一差距反映了噪声的影响，以及含噪 模型所带来的额外鲁棒性。

值得注意的是，含噪 模型与干净 模型的精确率相当，但后者的召回率较低。这可能是由数据分布偏移造成的：干净 模型是在干净词元上训练的，却是在含噪文本上评估的。模型在测试时首次遇到受扰动词元，便会变得保守并跳过边界跨度，导致更多假负例，从而降低召回率。

人工 模型（这是一个很出色的结果）优于含噪 模型，表明注入合成数据甚至可以提高原始准确率。

如第4.1节所述，为了保持可比性，本次实验评估仅通过噪声注入适度扩大了训练集规模。在实际应用中，可以通过 NAT 大幅扩充数据集，预计还能进一步提升人工 模型的表现。

所述的基本排序——人工 模型表现最佳，其次是含噪 模型，最后是干净 模型——在多次评估中始终保持一致；这些评估使用了不同的随机种子来进行数据集划分和模型初始化。

5.3 各实体的准确率

表5列出了使用F_1 分数评估的各实体表现，采用的是人工 模型。如表2所示，已标注实体的数量不同，导致各实体的识别率也有所不同。此外，有些实体本身就更

表 4：各模型的准确率。

模型	精确率	召回率	F_1 分数
含噪	73.4 %	82.0 %	77.5 %
干净	73.5 %	76.9 %	75.2%
人工	75.7%	80.3%	77.9%

表5：实体准确率。

实体	F_1 分数
职位名称	87.9 %
职位类别	76.5 %
技能	38.5 %
学科	58.8 %
活动	40.0 %

难以检测。职位名称 的识别准确率最高。

技能
对于技能而言，除了标注实体数量较少之外，还存在第二个挑战：标注的职业教育与培训文件中的技能，在内容上不同于额外训练数据中的技能。职业教育与培训数据集中的培训条例包含从事某种职业所需的前置技能，包括身体和心理方面的技能；例如“手指灵巧度”（灵巧度）。相比之下，作为额外训练数据使用的 ESCO 数据主要包含在训练期间需要学习的可习得技能（例如编程语言）。因此，SKILL 的识别率相对较低。

这凸显了多阶段微调的一个核心挑战：当不同数据集对某个实体采用的标注规范存在差异时，会产生问题。在未来工作中，我们计划不再使用单一的总体 SKILL 实体，而是更细致地区分前置技能和可习得技能。

学科与活动
对于学科和活动，不存在额外训练数据（因此也不存在来自中间模型的预训练权重）。因此，这两个实体的识别率都较低。将F_1 分数与学科和活动的分数同职位名称的分数进行比较，可以清楚地看出多阶段微调的影响；后者受益于大量源自 KldB 的额外训练数据。

5.4 实体混淆

图5展示了使用含噪模型进行实体识别的混淆矩阵（词元级混淆；不含“O”标签）。

混淆矩阵表明，在区分特定实体方面存在一些困难：

对于职位名称与职位名称组，某个术语究竟

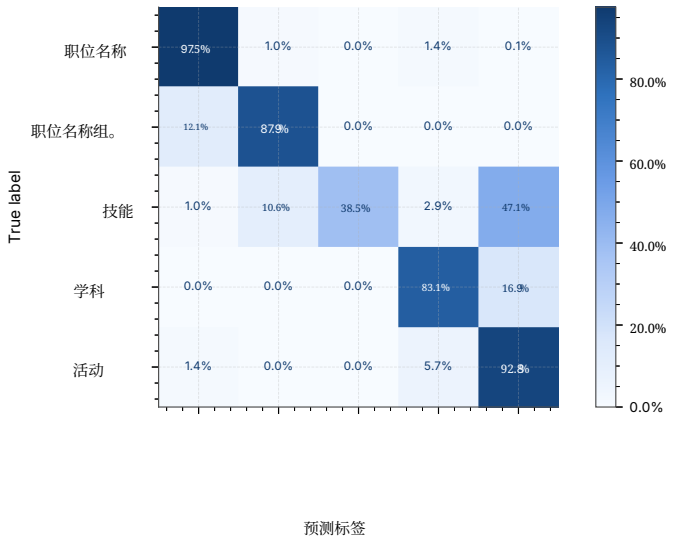

图5：实体识别的混淆矩阵。

指的是某种具体职业，还是更高层级的类别，并不总是明确；例如，“Fachar- beiter”（技术工人）。

对于技能 与活动，挑战在于，正如第5.3节所述，ESCO主要包含可学习的 技能，这些技能与工作活动相似。

对于活动 与学科，混淆风险较高，因为学科是课程主题的上位类别，而课程主题可能包含多项活动。

6 结论

本出版物研究了识别德国历史职业教育与培训文档中命名实体这一复杂任务，并着重指出了其中的上下文挑战。我们特别关注会传播至下游任务（如命名实体识别）的OCR错误。我们提出了一种针对职业教育与培训文档的稳健噪声感知命名实体识别方法，其表现优于基线模型和领域专用模型。我们训练并比较了三个模型：含噪、干净和人工，分别对应未经校正的OCR输出、经校正的OCR输出和合成扰动文本。人工 模型取得了最佳结果（F_1 分数：77.9%）。

我们进一步比较了不同实体类型的准确率。对于职位名称，其识别率最高（F_1 分数：87.9 %）。在定性分析中，我们还考察了识别不同实体类型时出现的混淆和固有难点。

总体而言，结果证明噪声感知训练对历史职业教育与培训文件有效。

6.1 局限性与未来工作

这些结果推动我们继续开发面向职业教育与培训文件的复杂文档处理流水线，尤其是整合更多预处理方法、版面分析和表格识别技术，以提升识别质量。

OCR错误

OCR引擎（如 Tesseract）通常内置预处理方法，例如二值化和倾斜校正。

BIBB 保存的历史职业教育与培训文档均经过专业数字化处理，因此扫描时光照不均并非问题。然而，由于年代久远，许多文档的纸张已经泛黄，文字也逐渐褪色。这些情况可能导致标准二值化方法失效，因此需要采用更先进的局部自适应二值化方法。

倾斜校正也是如此：在背景含噪、对比度低或旋转角度较大的情况下，标准方法可能会失效。

因此，未来的一项工作是在进行光学字符识别之前实施定制预处理方法，并采用后校正策略来提高光学字符识别输出的质量。视觉语言模型（VLMs），例如 LayoutLMv3（Huang 等人，2022年），也有望用于融入基于版面的特征。

本研究使用了最先进的 OCR 引擎 Tesseract；未来工作的一个开放方向是评估我们的研究结果在不同光学字符识别系统中的稳定性。

复杂页面布局

复杂页面布局是命名实体识别面临的另一项关键挑战。在某些情况下，由于换行，实体无法被识别，而此时阅读顺序也无法得到正确确定。因此，下一个重要目标是在命名实体识别之前，将版面分析和表格识别整合到文档处理流水线中，以处理复杂页面布局。

从文档中自动抽取职业信息仍然是一个重要问题。在本出版物中，我们为职业教育与培训文件的分析建立了一个可扩展处理流水线的基础。

作者贡献： 概念化：A.E.；数据整理：A.E.；方法论：A.E.；写作、审阅与编辑：A.E.、J.D.；调查：J.D.；可视化：A.E.；监督：J.D。

所有超链接最后一次核验日期为2025年9月19日。

参考文献

Akbik, A.、Blythe, D. 和 Vollgraf, R.（2018）。用于序列标注的上下文字符串嵌入。载于国际计算语言学会议（*COLING*），第 1638–1649 页。计算语言学协会。Alzahrani, E. 和 Jololian, L.（2021）。使用 BERT 模型的不同文本预处理技术如何影响作者的性别画像。*arXiv* 预印本 *arXiv:2109.13890*。Baldwin, T. 和 Li, Y.（2015）。深入分析文本规范化对社交媒体的影响。载于*NAACL-HLT*，第 420–429 页。Bikel, D. M.、Schwartz, R. 和 Weischedel, R. M.（1999）。一种能够学习姓名中包含何种信息的算法。机器学习，34(1):211–231。Bojanowski, P.、Grave, E.、Joulin, A. 和 Mikolov, T.（2017）。利用子词信息丰富词向量。计算语言学协会汇刊，5:135–146。Boubt, M.、Zayane, M. 和 Hamadou, A.（2022）。置信度感知的神经网络 OCR 后校正。载于国际文档分析与识别会议（*ICDAR*），第 81–95 页。德国联邦就业局（2020）。2010年职业分类——2020年修订版，第 1 卷：系统分类与字母顺序分类及说明。Chen, X.、Wang, S.、Liu, Y. 和 Sun, M.（2021）。用于文本纠错的去噪序列到序列预训练。载于计算语言学协会研究成果（*ACL Findings*），第 462–474 页。Decorte, J.-J.、Van Hautte, J.、Demeester, T. 和 Develder, C.（2021）。JobBERT：通过技能理解职位名称。*arXiv* 预印本 *arXiv:2109.09605*。Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K.（2019）。BERT：面向语言理解的深度双向变换器预训练。载于*NAACL-HLT*，第 4171–4186 页。计算语言学协会。Dobischat, R.、Käpplinger, B.、Molzberger, G. 和 Münk, D.（2019）。数字化及其影响：炒作还是革命？教育 *2.1 f¨ur Arbeit 4.0*？，第 9–24 页。Dorau, R. 和 Hein, K.（2025）。迈向依据 KldB 对德语职位名称进行自动分类。载于计算机科学与信息系统会议（*FedCSIS*）。Dörpinghaus, J.、Binnewitt, J.、Winnige, S.、Hein, K. 和 Krüger, K.（2023）。迈向德国劳动力市场本体：挑战与应用。应用本体，18(4):343–365。

Eftimov, T., Korou˘si´c Seljak, B., and Koro˘sec, P. (2017). 一种基于规则的命名实体识别方法，用于从循证膳食建议中进行知识抽取。*PLOS ONE*，12(6)。欧盟委员会（2025）。ESCO：欧洲技能、能力和职业数据库。Fischer, A. and D¨orpinghaus, J. (2024). 对在线资源进行网络挖掘，用于德国劳动力市场研究与教育：寻找事实真相? *Knowledge(MDPI)*，4(1):51–67。Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). 学习157种语言的词向量。*arXivpreprint arXiv:1802.06893*。Gururangan, S., Marasovi´c, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). 预训练不应止步：使语言模型适应领域和任务。*arXivpreprint arXiv:2004.10964*。Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., and Steiner, S. (2017). 基于统计学习的三种职业编码方法。*JournalofOfficialStatistics(SAGE)*，33(1):101–122。Hartig, J. (2008). 能力作为教育过程的结果。载于职业教育领域中的能力评估¨：理论、概念与方法，第26卷，第15–26页。德国联邦教育与研究部，柏林。Helmrich, R., Tiemann, M., Troltsch, K., Lukowski, F., Neuber-Pohl, C., Lewalder, A. C., and G¨unt¨urk-Kuhl, B. (2016). 工作场景数字化：劳动世界并未两极分化，但结构转型和工作岗位转换加速。第180号。WDP。Hendrycks, D., Lee, K., and Mazeika, M. (2019). 使用预训练可以提升模型鲁棒性并降低不确定性。载于国际机器学习大会（*ICML*），第2712–2721页。HR-XML 联盟（2007）。HR-XML 能力架构。访问日期: 2025-01-15。Huang, W., Rogers, A., and Rudinger, R. (2020). Skill2Vec：为职位匹配嵌入技能。载于经济学与自然语言处理 *EMNLP* 研讨会，第111–121页。Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. (2022). LayoutLMv3：通过统一的文本与图像掩码进行文档人工智能预训练。Huang, Z., Xu, W., and Yu, K. (2015). 用于序列标注的双向 LSTM-CRF 模型。*arXivpreprint arXiv:1508.01991*。IEEE 学习技术标准委员会（2008）。IEEE 可重用能力定义标准。载于*IEEE Std 1484.20.1-2007*。Jurafsky, D. and Martin, J. H. (2025). 《语音与语言处理：自然语言处理、计算语言学与语音识别导论》。

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. (2022). 无 OCR 文档理解 Transformer。载于欧洲计算机视觉会议，第498–517页。Springer。Konya, I. (2013). 用于鲁棒文档图像理解的自适应方法。博士论文，波恩莱茵弗里德里希·威廉大学。Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). 条件随机场：用于分割和标注序列数据的概率模型。载于国际机器学习大会（*ICML*），第282–289页。Morgan Kaufmann 出版公司。Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). 用于命名实体识别的神经网络架构。载于*NAACL-HLT*，第260–270页。计算语言学协会。Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2021). TrOCR：基于 Transformer 的光学字符识别与预训练模型。Li, N., Kang, B., and De Bie, T. (2023). SkillGPT：一种利用大语言模型进行技能抽取与标准化的 RESTful API 服务。*arXivpreprint arXiv:2304.11060*。Linten, M. and Pr¨ustel, S. (2015). 职业教育中的能力：概念、获取、评估与测量。载于选编自：职业教育文献数据库。BIBB，波恩。Mayhew, S., Gupta, N., and Roth, D. (2020). 通过真实大小写预训练实现鲁棒命名实体识别。载于*AAAI* 人工智能会议，第34卷，第8480–8487页。McCallum, A. and Li, W. (2003). 使用条件随机场、特征归纳和网络增强词典进行命名实体识别的早期成果。载于*HLT-NAACL*，第188–191页。Nadeau, D. and Sekine, S. (2007). 命名实体识别与分类综述。*LingvisticaeInvestigationes*，30(1):3–26。Namysł, M. (2023). 非结构化文档的鲁棒信息抽取。博士论文，波恩莱茵弗里德里希·威廉大学。Namysł, M., Behnke, S., and K¨ohler, J. (2021). 经验误差建模可提升含噪神经序列标注的鲁棒性。载于计算语言学协会研究成果（*ACL-IJCNLP*），第314–329页。计算语言学协会。Namysł, M., Behnke, S., and K¨ohler, J. (2020). NAT：用于鲁棒神经序列标注的噪声感知训练。载于计算语言学协会年会（*ACL*），第1501–1517页。计算语言学协会。

Namysł, M. 和 Konya, I.（2019）。使用深度学习实现高效的无词典光学字符识别（OCR）。载于国际文档分析与识别会议（*ICDAR*），第 295–301 页。IEEE。国家 O*NET 开发中心（2020）。O*NET 资源中心。Nguyen, K. C.、Zhang, M.、Montariol, S. 和 Bosselut, A.（2024）。使用大语言模型重新思考就业市场领域中的技能抽取。*arXiv* 预印本 *arXiv:2402.03832*。Paszke, A.、Gross, S.、Massa, F.、Lerer, A.、Bradbury, J.、Chanan, G.、Killeen, T.、Lin, Z.、Gimelshein, N.、Antiga, L. 等（2019）。PyTorch：一种命令式、高性能的深度学习库。神经信息处理系统进展（*NeurIPS*），32。Pennington, J.、Socher, R. 和 Manning, C.（2014）。GloVe：用于词表示的全局向量。载于自然语言处理经验方法会议（*EMNLP*），第 1532–1543 页。计算语言学协会。Rabiner, L. R.（1989）。关于隐马尔可夫模型及其在语音识别中选定应用的教程。*IEEE* 汇刊，77(2)：257–286。Reiser, T.、D¨orpinghaus, J.、Steiner, P. 和 Tiemann, M.（2024）。迈向数字化历史德国职业教育与继续职业教育法规数据集。《数据》，9(11)。Reiser, T.、D¨orpinghaus, J. 和 Tiemann, M.（2025）。检测德语文本中的职业：挑战与数据。载于人工智能与社会、教育及教育研究国际研讨会（*AISEER*）。Safikhani, P.、Avetisyan, H.、F¨oste-Eggers, D. 和 Broneske, D.（2023）。借助层级特征实现自动化职业编码：一种以数据为中心的预训练语言模型分类方法。《探索人工智能》，3(1)：6。Saluja, A.、Anderson, A. 和 Mofrad, P.（2021）。使用 T5 的自监督文本纠错。载于*NAACL*。Schierholz, M.（2014）。职业调查编码自动化。硕士论文，慕尼黑大学。Schierholz, M. 和 Schonlau, M.（2021）。用于职业编码的机器学习——一项比较研究。《调查统计与方法学杂志》（*OUP*），9(5)：1013–1034。Schiersmann, C.（2022）。国家继续教育战略背景下的继续教育咨询：财务与结构方面。《黑森州¨教育¨通讯》，72(1)：43–53。Singh, M.、Shashidhar, N. 和 Odongo, J.（2019）。使用 BERT 进行 OCR 错误纠正。载于*IEEE* 国际大数据会议（大数据），第 4032–4037 页。

Smith, R.（2007）。Tesseract 光学字符识别引擎概述。载于国际文档分析与识别会议（*ICDAR*），第 2 卷，第 629–633 页。Tabares, A.、Corchuelo, R. 和 Ruiz, E.（2018）。教育领域构建基于能力的本体的方法。《基于知识的系统》，145：1–14。Tiemann, M.、D¨orpinghaus, J. 和 Shivashankar, V. H.（2024）。采用双语本体方法对 OJAs 中计算机科学的分析。载于*INFORMATIK 2024*，第 2047–2058 页。德国信息学学会 e.V.。Tjong Kim Sang, E. F. 和 De Meulder, F.（2003）。CoNLL-2003 共享任务简介：与语言无关的命名实体识别。载于*HLT-NAACL*，第 142–147 页。Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, A. N.、Kaiser, Ł. 和 Polosukhin, I.（2017）。注意力机制就是你所需要的一切。神经信息处理系统进展（*NeurIPS*）。Viterbi, A.（1967）。卷积码的误差界及渐近最优译码算法。《*IEEE* 信息论汇刊》，13(2)：260–269。Wang, A. 和 Russakovsky, O.（2023）。使用微调数据覆盖预训练偏差。载于国际计算机视觉会议（*ICCV*），第 3957–3968 页。Wang, J. M. 和 Sun, Z.（2025）。从招聘广告中进行大语言模型监督的多语言技能抽取与分类。载于自然语言在信息系统中的应用国际会议（*NLDB*），第 94–104 页。Weinert, F. E.（2001）。学校中的比较性绩效测量——一种备受争议的不言自明之理。载于学校中的绩效测量，第 17–32 页。Beltz。Wolf, T.、Debut, L.、Sanh, V.、Chaumond, J.、Delangue, C.、Moi, A.、Cistac, P.、Rault, T.、Louf, R.、Funtowicz, M.、Davison, J.、Shleifer, S.、von Platen, P. 等（2022）。Transformer。Xu, G.、Ding, W.、Fu, W.、Wu, Z. 和 Liu, Z.（2021）。通过多源噪声模拟和困难样本挖掘实现稳健的文本分类学习。载于机器学习与数据库知识发现联合欧洲会议（*ECML PKDD*），第 285–301 页。