20 分钟-XD：瑞士新闻文章的可比语料库

Michelle Wastl[1] **Jannis Vamvas**[1] **Selena Calleri**[2] **Rico Sennrich**[1]
[1]Department of Computational Linguistics, University of Zurich [2]20 Minuten (TX Group)
{wastl,vamvas,sennrich}@cl.uzh.ch, {selena.calleri}@20minuten.ch

Abstract

我们介绍了 *20 分钟-XD*(**20 分钟** uten **交叉**-lingual**d**ocument-level)，这是一个源自瑞士在线新闻媒体 *20 分钟*/*20 分钟*的法语-德语文档级别的可比语料库。我们的数据集包含大约 15,000 对文章，时间跨度从 2015 年到 2024 年，基于语义相似性自动对齐。我们详细描述了数据收集过程和对齐方法论。此外，我们还提供了对该语料库的定性和定量分析。最终的数据集展示了广泛的跨语言相似度范围，从近乎翻译的文章到松散相关的文章，使其对于各种自然语言处理应用和广泛的语言学研究具有价值。我们公开发布了文档对齐和句子对齐版本的数据集以及描述的实验 [1,2] 的代码。

1 介绍

跨语言数据集在自然语言处理（NLP）中扮演着关键角色，支持包括双语文本挖掘、机器翻译和跨语言信息检索等一系列任务。其中，平行语料库——包含各语言中内容相关但非完全相同文本对的数据集——特别有价值。与由直接翻译组成的平行语料库不同，对比语料库自然地包含了精确翻译、释义以及松散相关内容的混合，反映了语言和文化之间的差异。这使它们成为训练和评估多语言 NLP 模型 (Lewis et al., 2020; Liu et al., 2020; Philippy et al., 2025) 的重要资源。

然而，现有的文档级跨语言语料库范围仍然有限。许多可用的资源以英语为中心，主要涵盖英语和另一种高资源语言，并且/或者仅限于句子级别的对齐而不是完整的文档 (Zweigenbaum et al., 2017; Artetxe and Schwenk, 2019)。与此同时，大型语言模型（LLMs）和现代化编码器架构在处理更长文本和多种语言方面的能力正在提高，进一步增加了对多/跨语言、文档级语料库的需求 (Hengle et al., 2024; Wang et al., 2024; Zhang et al., 2024)。

除了其在自然语言处理方面的应用外，跨语言文档级数据集还促进了更多以语言学为基础的研究，如跨文化话语分析 (Carbaugh and Cerulli, 2017) 或比较新闻研究 (Hanitzsch, 2019)。更具体地说，一个德法新闻文章语料库可以用来考察德国和法语地区新闻叙事和框架策略的差异。

鉴于这些潜在的跨学科应用场景，我们从在线瑞士新闻媒体 *20 分钟*/*20 分钟*收集了德语和法语的可比新闻文章。由于这两个版本由同一出版商制作，并且有一个内部的文章转换流程，从一种语言到另一种语言，它们在主题上有很高的重叠度，非常适合创建可比语料库。我们的数据集包含 15,000 对文章，跨越近十年（2015 – 2024）。每篇文章对由同一天发布的德语和法语文章组成，报道相同的或高度相关的事件。除了文档级别的对齐之外，我们还发布了该数据集的句子对齐版本，其中包含每个语言 117,126 个句子。

[1]数据集: https://huggingface.co/datasets/ZurichNLP/20min-XD
[2]代码: https://github.com/ZurichNLP/20min-XD

我们向研究社区发布该数据集用于非商业性的科学研究目的[3]。

2 相关工作

瑞士的多语言环境，拥有四种官方语言，为跨语言语料库的创建提供了肥沃的土地。几项先前的研究利用了这种语言多样性来构建多语言数据集。例如，SwissAdmin (Scherrer et al., 2014) 是一个由官方瑞士政府新闻稿组成的句子对齐语料库，提供德语、法语、意大利语和英语版本。同样地，Bulletin Corpus (Volk et al., 2016) 将瑞士信贷通讯公报的各个问题在同样的四种语言中进行了对齐。

20 分钟 也曾作为以前的自然语言处理相关研究的资源。Rios et al. (2021) 构建了一个自动文本简化的数据集，通过将原始德语 *20 分钟*文章与其简化版本配对来实现这一点。最近，Kew et al. (2023) 创建了一个旨在自动摘要德语新闻数据集，进一步扩展了瑞士新闻数据在自然语言处理研究中的应用。

通过这项工作，我们希望通过引入 *20 分钟-XD*，一个源自 *20 分钟*（德语）和 *20 分钟*（法语）的文档级可比语料库，来连接这两个主题。

3 数据采集

为了构建我们的数据集，我们首先从 `www.20min.ch/` 和 `www.20min.ch/fr/` 爬取了总计 593,897 篇在线新闻文章，涵盖了从 2015.01.01 到 2024.12.01 的时间段。在以下子章节中，我们将描述用于识别和对齐语义相关文章的过程。

3.1 验证集

为了建立对齐评估的金标准，我们选择了来自单一出版日的所有文章，结果得到了 87 篇德语文章和 70 篇法语文章。每篇法语文章都被手动与德语文章进行比较以识别可比配对。虽然我们没有严格禁止 n:n 配对，但最终的验证集只包含 1:1 配对。通过这一过程，我们将 28 篇文章对齐为 14 对，形成了我们的验证集。详细统计信息请见表 1。

[3]参见附录 A 以获取详细的版权通知。

3.2 自动文章对齐

由于手动对齐跨语言的可比文章耗时且需要精通德语和法语，我们利用多语言嵌入模型自动化这一过程。具体来说，我们将每篇文章的部分内容编码为数值向量，并计算余弦相似度得分，该得分范围从 -1 到 1 (∗100)，以量化它们的语义相似性。

为了找到最适合 *20 分钟*篇文章的对齐方法，我们在验证集上使用不同的嵌入模型、对齐方法和相似度阈值进行了实验。

我们选择不嵌入全文内容，以确保在测试模型之间进行公平比较，其中一些模型具有序列长度限制（5 个测试模型中有 3 个）。我们在验证集上的结果表明，连接文章的标题和开头部分提供了足够强的文档对齐信号。这使得可以使用基于编码器的嵌入模型进行资源高效实验，同时避免了长度限制。

3.2.1 模型

我们实验了表 2 中的一组模型：`多语言改写-mpnet` 是一种最先进的多语言句子级释义识别模型 (Song et al., 2020)；`gte 多语言基础版本`，一种长上下文多语文本表示模型 (Zhang et al., 2024)；`句子-瑞士 BERT`，基于 sentence-BERT 的模型 (Reimers and Gurevych, 2019)，在领域内数据 *20 分钟*上训练的 (Grosjean and Vamvas, 2024)；`gte-modernbert-base`，现代版、更高效、长上下文版本的 BERT，主要在英语数据上进行了训练 (Warner et al., 2024)。

初步实验表明，基于大型语言模型 (Wang et al., 2024) 的模型在性能上超过了基于编码器的模型，并且能够处理更长的输入序列。然而，它们也增加了嵌入过程的计算复杂性，使其在扩展到更大数量文档时变得非常消耗资源，在内存和时间方面几乎不可行。

	验证集		完整数据集		前 1.5 万	
统计	**德国**	**法国**	**德国**	**法国**	**德国**	**法国**
Total # of aligned articles	14	14	73,085	73,085	15,000	15,000
Total # of sentences	401	358	1,888,323	1,608,497	357,071	327,628
Total # of tokens	9,087	9,690	43,559,153	43,256,366	8,378,874	8,956,116
Total # of characters	38,523	38,519	189,598,932	174,789,207	36,924,383	36,387,070
Avg. title length in characters	59	54	51	53	51	54
Avg. title length in tokens	18	18	15	17	15	17
Avg. lead length in characters	146	155	152	146	152	150
Avg. lead length in tokens	39	43	39	40	38	41
Avg. content length in characters	2,547	2,542	2,391	2,192	2,258	2,222
Avg. content length in tokens	706	753	650	649	612	655
Avg. content length in sentences	29	26	26	22	24	22

表 1: 验证、完整和前 15k 子集的详细统计。句子分割使用了 spaCy '[de/fr]_core_news_sm' (Honnibal and Montani, 2017) 模型进行句段划分和分词，采用了多语言 `mppnet.paraphrase` 分词器。

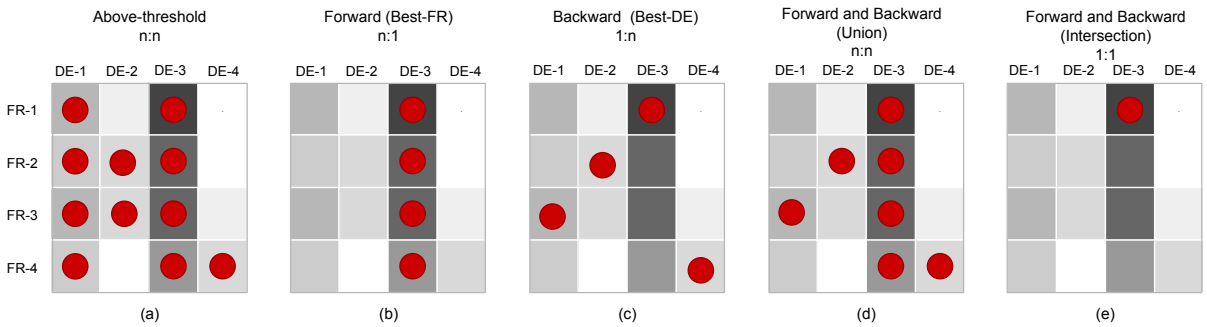

图 1: 不同对齐策略的矩阵可视化。

模型	**阈值以上**	**交集**	**并集**	**最佳-DE**	**最佳-FR**	**平均值**
多语言 mppnet 改写模型	54.1	**64.7**	54.1	57.8	55.8	57.3
gte 多语言基础版本	55.6	62.1	55.6	60.0	58.5	58.3
LaBSE	53.3	48.5	56.5	60.0	46.2	52.9
句子-瑞士 BERT	62.9	62.5	62.9	61.1	62.5	62.4
gte-modernbert-base	45.5	53.3	50.0	54.1	50.0	50.6

表 2: 不同模型和不同对齐方法在验证集上的 F1 性能比较。相应的阈值见附录 B。

3.2.2 对齐策略

先前的跨语言对齐工作考虑了多种可能的对齐策略，这些策略会根据不同的类别（如 Jalili Sabet et al. (2020) 中所述）扩展或限制最终的对齐数量。与 Hämmerl et al. (2024) 类似，我们实验了从弱到强的不同对齐策略。其中，较弱的对齐策略通常允许更高的语义相似度范围和多种可能的对齐方式，而较强的对齐策略则更倾向于高语义相似度，并且可能只包括一种好的对齐（图 1）。

阈值以上 认为所有相似性评分超过某个阈值的文档对都是可对齐的，允许进行多对多（n:n）

对齐。这意味着可以在不添加超出相似性阈值之外的其他约束条件下，将任意数量的法语文章链接到任意数量的德语文章。虽然这种方法捕捉到了潜在对齐的广泛范围，但它并不强制执行唯一性或最佳匹配约束，导致对齐的数量增加（图 1a）。

最佳-FR 应用了一对多（n:1，德语：法语）约束条件，其中每篇 FR 文章与它具有最高余弦相似度的单一 DE 文章对齐，前提是该相似度超过阈值。这确保了每篇 FR 文档都有一个最佳匹配的 DE 对应物，但多篇法语文章仍然可以映射到同一篇德语文章。这种方法优先考虑法语文章选择它们最近似的德语文本，同时允许不对称的对齐（图 1b）。

最佳-DE 遵循与 Best-FR 相同的原理，但从德语的角度出发，强制执行一对一多（1:n，德语：法语）的约束。这导致了一个设置，在此设置中，单个德语文章可以链接到多个法语文章，捕捉到一个德语文档是多个法语文档的最佳翻译候选者的场景（图 1c）。

并集 取 Best-DE 和 Best-FR 对齐的并集，允许多对多 (n:n) 对齐，但比 Above-Threshold 方法更为严格。它不再考虑所有阈值以上的成对组合，仅保留至少有一方将另一方选为其最相似文档且超过阈值的文档对（图 1d）。

交集 是最严格的策略，强制执行一对一（1:1）约束。只有当法语文章是德语文章的最佳匹配且反之亦然，并且它们的相似度分数超过阈值时，才会发生有效的对齐。该方法形成 Best-DE 和 Best-FR 的交集，确保对齐是双向的并且相互最优（图 1e）。

3.3 设置阈值

由于并非每篇文章在另一种语言中都有可比较的对应文章，因此我们定义了一个相似度得分阈值，超过该阈值则认为两篇文章是可以对齐的。这个阈值必须超过上述描述的所有对齐策略中的一个。为了确定最优阈值 θ，我们在 0 到 100 的范围内以 0.5 为步长进行迭代，选择能够最大化我们验证集上的 F1 分数的那个。

$$\hat{\theta} = \arg\max_{\theta \in \{0, 0.5, \ldots, 100\}} F_1(\theta)$$

我们定义 F1 如下，其中 P 表示预测的配对，G 表示黄金配对：

$$Prec = \frac{|P \cap G|}{|P|}$$

$$Recall = \frac{|P \cap G|}{|G|}$$

$$F_1 = 2 \cdot \frac{Prec \cdot Recall}{Prec + Recall}$$

此过程针对上述描述的每个嵌入模型重复进行。我们的结果显示，使用对齐策略交集的多语言改写-mpnet 在相似度分数阈值为 46 时，在验证集上优于所有其他模型（见表 2），使其成为我们文章对齐的方法。值得注意的是，我们的验证集样本数量较少（87 篇德语和 70 篇法语文章）。这可能导致统计噪声，夸大结果中明显的差异，使它们看起来比实际情况更大/更小。

3.4 选择时间窗口

为了确保精确对齐并减少计算复杂性，我们将比较限制在同一天发表的文章。这种方法最小化了讨论类似主题但具体事件或发展上无关的文章之间的虚假匹配。

3.5 后处理

对齐法语和德语文本后，我们清理得到的语料库。人工检查表明，有问题的文章通常具有可疑的高相似度分数，并包含错误消息或相同语言中的相同文本。我们删除这样的配对。

3.6 句子对齐

为了对数据集提供更细致的洞察，我们进行了句子级别的分析。为此，我们首先使用 spaCy 的 '[de/fr]_core_news_sm' (Honnibal and

Montani, 2017) 模型将文章分割成句子，用于德语和法语。

一旦分段后，我们执行跨语言句子对齐，再次应用上述描述的最佳方法：*多语言改写-跨语种_mpnet* 以及*交集*对齐策略。虽然我们在分析中仅考虑相似度分数高于 46 的句子对，但我们发布的是语料库中的所有对齐句子版本，包括那些相似度分数未达到阈值的句子。这允许进行更全面的未来分析，不仅捕捉最强烈的对齐句子，还包括那些仍可检测到最弱语义相似性的句子。

我们对数据集的句子级版本进行后处理，移除包含少于 30 个字符的句子对，这包括名称、尾随字符和源缩略词。

3.7 相似性的其他度量方法

在第 4 节的语料库描述中，除了余弦距离外，我们还利用基于句子对齐的其他跨语言相似性度量。

每篇文档中可对齐的句子数量 为了估计文章中相似文本的比例，我们计算可对齐句子的相对百分比。这一度量特别有趣，因为在自动文章对齐过程中并未考虑整篇文档，如子章节 3.2 所述。对于每篇文章，我们将可对齐句子比率定义为：

$$\text{AlignRatio} = \frac{\text{NumAlignedSentences}}{\text{TotalSentences}}$$

句子长度相关性 如果句子长度，以句子中的字符数来衡量，在两种语言之间存在系统性差异，则对齐文章中句子长度之间的高相关性可以作为语义相似性的额外指标。因此，我们计算一篇文章的句子长度相关性为皮尔森相关系数。

单调性 我们通过计算对齐句子位置的肯德尔等级相关性来测量一对对齐文章之间的跨语言单调性（对齐句子以相同顺序出现的程度）。

4 数据集

我们的对齐过程产生了 74,507 篇文章配对。在后处理阶段，语料库被过滤到 73,085 篇文章配对。根据与 *20 分钟*的协议，我们发布的数据集限制为 30,000 篇文章。因此，我们选择了按相似度分数排序的前 15,000 篇文章配对进行发表，我们在以下内容中将其称为顶级 15k 数据集。尽管如此，在本文的其余部分，我们将考虑完整的数据集和前 15k 篇文章配对作为分析的对象。两者的详细数据集统计信息如表 1 所示。

在来自前 15k 数据集的每种语言共 300,000 多个句子中，我们对齐了每种语言的 133,693 个句子，在过滤后剩下 117,126 个。对于第 4.2 节的相关性研究，我们考虑所有相似度分数高于 46 的句子对，总计有 109,871 个句子对。

4.1 定性分析

表 3 提供了对顶级 15k 数据集中具有最低、平均和最高余弦相似度分数的文章对，以及所有 75,085 篇最初配对文章中相似度得分最低的文章对的定性比较。得分最高的文章对表现出强烈的词汇和句法相似性。平均得分的文章对有效传达了相同的意思，但在信息呈现顺序上显示出明显的差异。仅在德语导语的最后一句话以及法语导语的最后一短语中引入了不同的信息。顶级 15k 数据集中得分最低的一对文章涵盖了相同的事件，但在选词和信息传递的顺序上有显著差异。整个配对文章集中的得分最低的文章对虽然仍然松散相关（金融危机），但在实际描述的事件上有所不同（例如，导致公司崩溃的法庭案件与子公司斗争）。

这些结果表明，我们的数据集主要由涵盖同一主题但语义重叠程度、文本结构和长度不同的文章组成。为了进一步了解这些特征及其与语义相似性的关系，我们对对齐文章的余弦分数与第 3.7 节中描述的不同度量进行了相关性研究。

相似性分数	德国	法国
余弦：98.48 (**最大值**) 句子长度相关性：0.75 对齐比率 DE：0.68 对齐比率 FR：0.56 单调性：1.0	**标题：** 移动性。：ń自 2030 年起，我们仅提供全电动汽车ż **引言：** 电动革命正在到来。传统汽车制造商目前处境艰难。我们采访了 Helen Hu，Volvo 这个自 2010 年以来由中国控股的瑞士分支机构的首席执行官，询问她如何看待移动性的未来。	**标题：** 移动性：「从 2030 年起，我们将只提供全电动汽车」 **引言：** 电动革命正在进行。传统汽车制造商目前处境艰难。我们询问了自 2010 年以来由中国拥有的沃尔沃瑞士分公司的负责人 Helen Hu，她是如何看待未来的移动性的。
余弦：84.05 (**平均值在前 1.5 万中**) 句子长度相关性：-0.78 对齐比率 DE：0.23 对齐比率 FR：0.21 单调性：-1.0	**标题：** LKW 撞上了货车，然后坠落 **领导：** 一辆卡车在周二坠落了 300 米。66 岁的司机严重受伤。现在有了关于事故原因的初步发现。	**标题：** 一辆卡车从 300 米高处坠落，司机幸存 **领导：** 一名重型货车司机周二在车辆偏离道路后严重受伤，事件发生在乌里州。
余弦：78.65 (**前 1.5 万中的最小值**) 句子长度相关性：-0.47 对齐比率 DE：0.07 对齐比率 FR：0.2 单调性：-0.3	**标题：** 巴西大奖赛 - 博塔斯赢得冲刺赛 - 汉密尔顿疯狂追赶至第 5 名 **领导：** 在周六的巴西大奖赛中进行了冲刺赛决策。瓦尔特里 · 博塔斯确保了 3 个世界冠军积分和周日比赛的杆位。	**标题：** 汽车 – 缪尼克 · 维斯塔潘因冲刺赛失利而未能获得胜利和杆位 **领导：** 瓦尔特里 · 博塔斯赢得了冲刺赛，并将在周日的巴西大奖赛中从首位发车。马克斯 · 维斯塔潘将排在他身后，刘易斯 · 汉密尔顿则在第 10 位。
Cosine: 46.00 (**数据集中的最小值**)	**标题：** 他的孪生兄弟将他告上法庭 **引言：** 高风险的股市交易导致一家著名的楚尔信托公司破产，这些交易是由他们的监事会主席进行的。被告必须出庭。	**标题：** 被凯罗斯拖累，Julius Bär 需要弥补 **引言：** Julius Bär 的意大利子公司几乎看起来像是这家瑞士财富管理公司的所有问题的源头。

表 3: 获得最低、平均和最高余弦相似度分数的前 15,000 篇对齐文章以及从最终数据集中过滤出的整体得分最低的对齐文章的标题和导语文本的比较。

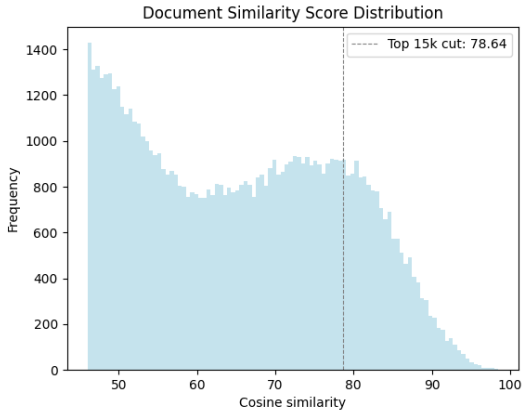

图 2: 文档（余弦）相似度得分分布覆盖了所有 74,085 篇文章对，并将其分为从阈值 46 到 100 的 100 个区间。虚线表示的是高于该切割点的前 15,000 篇文章对构成了最终可比较的数据集。

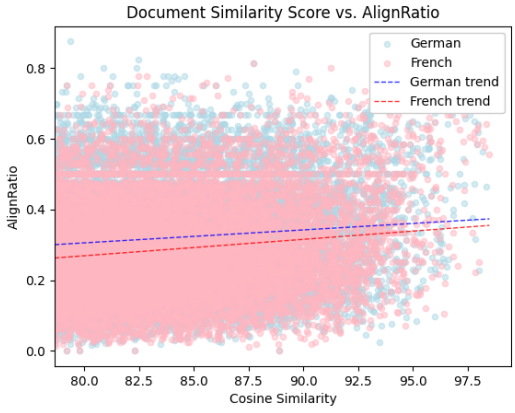

图 3: 文档余弦相似性与德语和法语中每篇对齐文章的 AlignRatio 相比。两种语言都显示出正向趋势线，具有弱正相关（法语：皮尔森相关系数 $r = 0.145$；德语：$r = 0.103$）。

4.2 定量分析

4.2.1 余弦相似度分布

图 2 展示了对齐文章中的余弦相似度分数分布。该分布显示出右偏斜模式，表明在抓取的文章集合中，具有中等语义相关性的法语和德语文章比那些具有极高相似性得分的文章更为普遍。随着余弦相似度的增加，文章数量首先下降然后再次上升，在大约 80 处达到一个小峰值，位置几乎正好处于我们的 15,000 个上限点。超过这个上限后，文章对的频率急剧下降

到相对较低水平，向更高的相似性得分靠近。这种模式大致表明存在两组文章对：一组代表中等相关度的文章，另一组不太明显，代表着更加密切相关的文章。

4.2.2 与对齐比率的相关性

作为一种进一步的语义相似度衡量标准，我们采用对齐比率（AlignRatio），它测量两种语言文章中对齐句子的比例，并考察文档相似度得分的相关性。如图 3 所示，德语和法语都表现出余弦相似度得分与对齐比率之间的弱正相关关系（$r = 0.145$ 用于法语，$r = 0.103$ 用于德语）。这些发现表明，在全文中具有更多对齐的文章倾向于有稍微更高的语义相似性。这支持了我们的假设，即仅依赖标题和导语进行自动对齐是足够的但并不完美。

4.2.3 与句子长度的相关性

为了分析文档相似度分数与对齐文章中句子长度变化之间的关系，我们计算了余弦相似度分数和每对文章的句子长度相关性之间的关联。如图 4 所示，结果表明存在非常弱的正相关（$r = 0.084$）。

4.2.4 与单调性的相关性

我们还研究了文档相似性得分与单调性的关系，后者量化了信息（=句子）顺序在对齐文章之间被保留的程度。图 5 展示了余弦相似性得分与单调性之间的相关性，显示出较弱的正相关（$r = 0.147$）。这表明，类似于之前的结论，虽然较高的文档相似性得分略微关联于信息的更单调对齐，但这种效果并不强。接近 -1.00 和 1.00 的聚类可能表示只有少数一两个对齐句子的文章数量很高——这一模式值得进一步调查。

鉴于我们的定性分析和相关性研究，我们相信我们的数据集保持了足够的质量以作为可比较的语料库，涵盖了直接翻译到相当不相关的文本序列之间的整个范围。然而，进一步使用这些指标可能会提供更多的见解。具体来说，

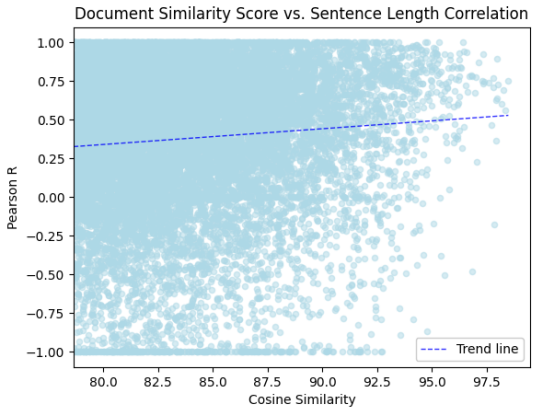

图 4: 文档的余弦相似度与每篇对齐文章的句子长度相关性的比较。在这两个变量之间可以检测到非常弱的正相关趋势（皮尔逊相关系数 $r = 0.084$）。

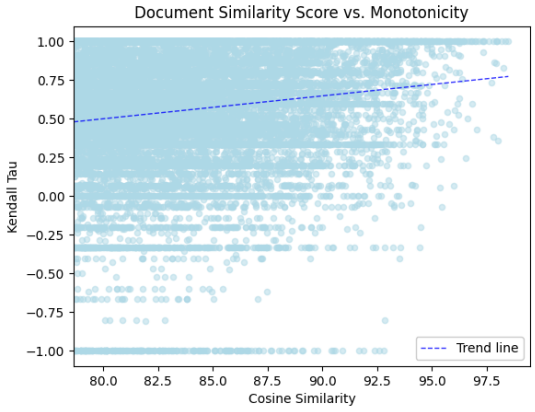

图 5: 文档的余弦相似性与每篇对齐文章的单调性分数相比。在这两个变量之间可以检测到一个弱正相关趋势（皮尔逊相关系数 $r = 0.147$）。

对齐比例可能作为一个指标来表明哪些信息在两个语言区域中被认为是重要的以及哪些信息缺失于其中一个或另一个。类似地，句子长度的相关性可以为新闻特定的翻译研究提供有价值的视角。最后，单调性可以通过分析主题特定的趋势进一步探索，可能会揭示出哪些主题倾向于被以更单调的方式进行翻译。

5 未来工作

5.1 比较全文的相似性

虽然仅使用标题和导语足以对齐相似的文章，但将全文内容纳入相似度分数计算可以提

供更细致且准确的语义相似性和相关性洞察。这种方法能够更细腻地表示叙事结构、论证以及主题强调。尽管计算量大，现代嵌入模型如 `e5-instruct-7b` 或 `gte` 多语言基础版本理论上可以处理较长的文字段落，使得全文比较变得越来越可行。

5.2 多语言长上下文嵌入模型

基于编码器的嵌入模型目前正在经历一场复兴，采用了现代化的实现方式，如 ModernBERT (Warner et al., 2024)，显著提高了效率并增强了处理更长文本序列的能力。目前，针对文本相似性任务指定的多语言版本模型十分稀缺。未来的工作可以探索将 ModernBERT 扩展到多语言环境和/或跨语言文档对齐的优化。另一个潜在方向是利用这些现代架构来开发一个文档级别的与 (sentence-)swissBERT 模型相对应的模型。

5.3 差异识别

虽然语义相似性一直是自然语言处理的主要关注点，但检测和量化文本之间的差异——尤其是在不同语言之间——是一个新兴的研究领域 (Vamvas and Sennrich, 2023)。受版本控制中基于差分操作的启发，这一任务可能对自然语言版本管理、协作文档编辑以及编辑工作流程产生影响。Vamvas and Sennrich 表明语义相似性数据集可以被重新用于差异检测，但必须进行合成修改以涵盖跨语言性和较长文本序列。

鉴于我们在数据集中观察到的变化范围（见第 4 节），接近翻译和松散相关文章的多样性，通过扩展带有细粒度注释的语料库——在段落、句子甚至词汇层面——可以促进自动跨语言差异识别的研究。

6 结论

我们引入了 *20 分钟-XD*，一个全新的法德新闻文章文档级可比数据集，源自瑞士报纸 *20 分钟*/*20 分钟*。该数据集包含 15,000 篇对齐的文章（或 117,126 个对齐的句子），这些文章在十年间发表。为了建立文档级和句子级的对齐，我们使用了一个多语言释义识别模型，在人工整理的验证集上的实验中该模型表现出色。定性和定量的结果表明，我们的语料库捕捉到了跨语言相似性的广泛范围，从近似翻译到更松散相关的文本对，这些文本仍然覆盖了相同的事件，并且具有不同程度的可对齐句子、文本长度和单调性。我们预计它将在未来跨越广泛的以语言学为动机的研究中得到应用。

致谢

这项工作由瑞士国家科学基金会资助（项目 InvestigaDiff；编号 10000503 用于 MW、JV 和 RS，以及项目 MUTAMUR；编号 213976 用于 RS）。我们真诚感谢 20 Minuten（TX 集团）的所有人对他们的支持，并使他们的数据能够被研究社区访问，特别感谢 Dean Cavelti 耐心的沟通。我们也感激 Unitectra，特别是 Peter Loch，提供了宝贵的法律指导。最后，我们要向苏黎世大学计算语言学系表示感谢，感谢他们带来的启发性讨论和指导，特别感谢 Sarah Ebling、Andrianos Michail、Patrick Haller 和 Anastassia Shaitarova。

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Donal Carbaugh and Tovar Cerulli. 2017. *Cultural Discourse Analysis*.

Juri Grosjean and Jannis Vamvas. 2024. Fine-tuning the SwissBERT encoder model for embedding sentences and documents. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 41–49, Chur, Switzerland. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Hanitzsch. 2019. Comparative Journalism Research.

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2024. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. *Preprint*, arXiv:2408.10151.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Tannon Kew, Marek Kostrzewa, and Sarah Ebling. 2023. 20 Minuten: A multi-task news summarisation dataset for German. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 1–13, Neuchatel, Switzerland. Association for Computational Linguistics.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems*, volume 33, pages 18470–18481. Curran Associates, Inc.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Yves Scherrer, Luka Nerima, Lorenza Russo, Maria Ivanova, and Eric Wehrli. 2014. SwissAdmin: A multilingual tagged parallel corpus of press releases. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1832–1836, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Jannis Vamvas and Rico Sennrich. 2023. Towards unsupervised recognition of token-level semantic differences in related documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13543–13552, Singapore. Association for Computational Linguistics.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67,

Vancouver, Canada. Association for Computational Linguistics.

A 版权说明

生成的数据集附有以下版权声明：

德国 / 德语（原始）：

l' 2025. TX Group AG / 20 分钟.

此数据集包含 TX 集团 AG/20 分钟的受版权保护的材料。它仅用于非商业科学研究目的。任何未经 TX 集团 AG/20 分钟明确许可的商业使用、复制或传播都是被禁止的。

英语 / 德语：

l' 2025. TX Group AG / 20 分钟.

该数据集包含来自 TX 集团 AG/20 分钟的受版权保护的材料。仅供非商业科学研究用途提供。未经 TX 集团 AG/20 分钟明确许可，禁止任何商业使用、复制或分发。

B 验证集上的实验

模型	阈值以上	交集	并集	最佳-DE	最佳-FR
多语言 Mpnet 基础版本 v2 重新表述模型	61.5	46.0	61.5	47.0	46.0
LaBSE	66.0	50.5	50.5	50.5	50.5
句子-瑞士 BERT	74.5	69.5	74.5	73.0	74.5
gte 多语言基础	65.0	65.0	65.0	60.0	56.0
gte-modernbert-base	66.0	66.0	66.0	66.0	63.0

表 4: 不同模型和对齐方法的最佳阈值。相应的 F1 分数见表 2。