arXiv:2305.01211v1 [cs.CL] 2 May 2023

MultiLegalSBD：多语言法律句子边界检测数据集

Tobias Brugger*
瑞士伯尔尼大学
tobias.brugger@students.unibe.ch

Matthias Stürmer
瑞士伯尔尼大学 瑞士伯尔尼应用科学大学
matthias.stuermer@unibe.ch

Joel Niklaus*
瑞士伯尔尼大学 瑞士伯尔尼应用科学大学 美国斯坦福大学
joel.niklaus@unibe.ch

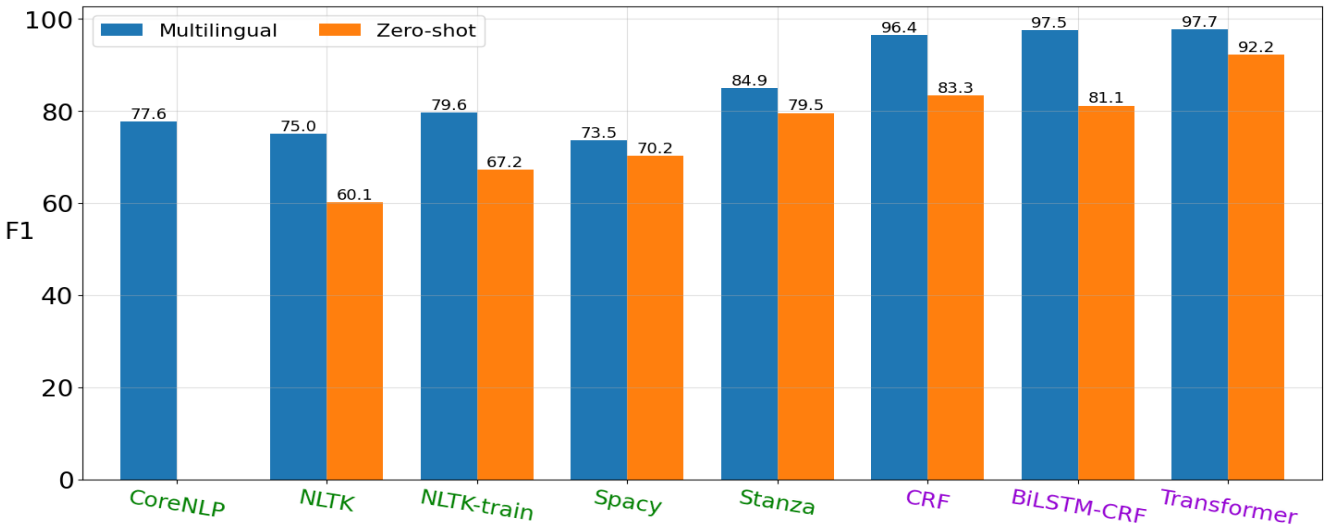

图 1：基线的平均性能比较（绿色）以及我们的多语言模型（紫色）在多语言法律文本（法语、意大利语、西班牙语、英语和德语）上的表现，以及在葡萄牙语数据上的零样本实验

摘要

句子边界检测（SBD）是自然语言处理（NLP）的基础组成部分之一，句子切分不正确会严重影响下游任务的输出质量。对于算法而言，这是一项具有挑战性的任务，尤其是在法律领域，因为该领域使用的句子结构复杂且各不相同。在本研究中，我们整理了一个多样化的多语言法律数据集，包含6种语言中超过130,000个经过标注的句子。我们的实验结果表明，现有句子边界检测模型在多语言法律数据上的性能不尽如人意。我们基于.....训练和测试了单语言模型和多语言模型。条件随机场、双向长短期记忆网络-条件随机场和 Transformer模型，展现了最先进的性能。我们还展示了，在葡萄牙语测试集的零样本设置下，我们的多语言模型优于所有基线。为鼓励社区进一步开展研究与开发，我们已公开提供数据集、模型和代码。

CCS 概念

• **应用计算** → 法律；标注；• **计算方法** → **自然语言处理**；监督学习；**机器学习**。

关键词

句子边界检测，自然语言处理，法律文档分析，文本标注，多语言

ACM 参考格式：
Tobias Brugger、Matthias Stürmer 和 Joel Niklaus。2023 年。MultiLegalSBD：多语言法律句子边界检测数据集。载于 第*19*届人工智能与法律国际会议（*ICAIL '23*）论文集。 ACM，纽约，纽约州，美国，10页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

*两位作者对本研究的贡献相同。

允许为个人或课堂使用而免费制作本作品的部分或全部数字或纸质副本，前提是这些副本不得用于盈利或商业用途，且副本必须附有本声明及首页上的完整引文。本作品中第三方组件的版权必须得到尊重。其他所有用途，请联系所有者/作者。*ICAIL '23*，*2023*年*6*月*19*日至*23*日，葡萄牙布拉加©2023 版权所有归所有者/作者所有。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM。
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

近期的方法学进展（例如 Transformer模型） [34], 推动了语言模型质量和性能的显著提升，也促进了一般自然语言处理（自然语言处理）领域的发展。这一趋势在法律自然语言处理中同样十分明显，近年来相关研究论文数量大幅增加 [14]。

句子边界检测（SBD）任务得到的关注和资源并不多，有些人认为该任务已经解决，因为通过采用简单的查找方法，捕捉句号、感叹号和问号等常见句末字符，并结合手工制定的规则 [26]，即可取得较高的基线性能。这种方法适用于新闻文章等格式规范且经过整理的文本。而对于噪声更多的领域特定数据，其文本结构各异，且许多句末字符存在歧义 [8, 15] ——例如，句号出现在缩写、省略号、首字母缩写等情况中时，并非句末标点——这类数据往往会使上述方法乃至更复杂的现成句子边界检测系统不堪重负。多个特定的句子边界检测应用都已证实这一点，例如用户生成内容 [9, 26]，以及临床领域 [20] 和金融领域 [7, 19]。

在法律文档中，上述困难因法律文本由段落、条款等较小部分组成而加剧，这使其与标准文本有很大不同。此外，句子较长，并且可能包含引文、括号和列表等复杂结构。这些结构通常用于向读者传达额外信息（例如，引用另一篇文本的引文），或以特定方式对文本进行排版（例如，列表能够突出观点或提高长段落的可读性）。然而，这些结构或特殊句子并不遵循标准句子结构，因此给句子边界检测系统带来了额外挑战；关于英语[27, 29]和德语 [10]法律文档的多项研究对此已有说明。

1.1 动机

拥有可靠的句子边界检测系统对于准确开展文本的自然语言处理分析至关重要。较差的句子边界检测可能导致错误传播到更高层次的文本处理任务中，从而阻碍整体性能。例如，多语言EUROPARL语料库的整理需要进行适当的句子边界检测，以便对齐两种语言中的句子，用于统计机器翻译。Koehn [16] 指出，句子边界检测存在困难，因为每种语言都需要专门的工具，而并非所有语言都能轻易获得这些工具。不充分的句子边界检测会削弱句子对齐算法的性能，并降低语料库的质量。因此，高质量的句子边界检测系统，尤其是针对法律领域定制的系统，能够显著提升性能。

另一个例子是否定范围解析（NSR），其重点是在句子中找出否定词（例如“not”）及其对周围词语含义的影响。否定词对于文本的语义表示至关重要，因为它们会反转命题的取值。这在法律领域尤其有用，能帮助从文档中提取信息的模型更好地理解输入文本的含义，例如根据确切措辞识别法院判决的结果。NSR 模型通常需要将数据拆分为句子，以便标注训练数据和应用输入，因此可靠的句子边界检测系统至关重要。句子边界检测系统对句子的错误判断可能会显著降低输入数据质量和模型性能。恰当的句子边界检测对于其他自然语言处理任务也至关重要，例如文本摘要、词性标注和命名实体识别，这些任务在法律领域都具有相关性。

1.2 主要研究问题

在这项工作中，我们提出并探讨三个主要研究问题：**RQ1:** 现有句子边界检测系统在法语、西班牙语、意大利语、英语和德语法律数据上的性能如何？**RQ2:** 通过基于条件随机场、双向长短期记忆网络-条件随机场和Transformer模型训练单语及多语言模型，我们能在多大程度上提升这一性能？**RQ3:** 多语言模型在未见过的葡萄牙语法律文本上的性能如何，即零样本实验的结果如何?

1.3 贡献

本文的贡献主要有两方面：

(1) 我们整理并公开发布了一个规模大、多样且高质量的多语言法律数据集（见第3节），其中包含超过 130,000 个经标注的句子跨度，供学界进一步研究。(2) 使用该数据集，我们展示了现有句子边界检测系统在法语、意大利语、西班牙语、英语和德语法律文本上的性能并不理想。我们训练并评估了基于条件随机场（CRF）、双向长短期记忆网络-条件随机场和Transformer模型的最先进单语句子边界检测模型，F1 分数最高达到 99.6%。我们展示了多语言句子边界检测模型的性能和可行性。这些模型在所有语言上进行训练，F1 分数达到较高的九十几，与我们针对上述各语言训练的单语模型相当或更优。在零样本实验中，我们通过在未见过的葡萄牙语法律文本上测试多语言模型，证明了实现良好跨语言迁移的可行性。我们公开发布了这些数据集[1]、所有单语和多语言模型[2]（见第5节），以及我们的代码[3]，供学界进一步使用。

2 相关工作

在本节中，我们将讨论现有文献。首先，我们将介绍展示为何需要对句子边界检测开展更多研究的相关工作。其次，我们将考察以多种语言解决法律文本中句子边界检测问题的相关研究。最后，为了内容的完整性，我们将探讨其他领域中的句子边界检测研究，并介绍法律领域的多语言数据集。

Read 等人 [26] 通过重新审视当前在英文新闻文章和用户生成内容上表现最先进的句子边界检测系统，对句子边界检测已被“解决”这一现状提出质疑，尤其是在更非正式的语言和特殊领域中。这些系统在前者上的 F1 分数达到了九十多分的高位；然而，在用户生成内容上的表现明显变差，分数降至九十多分的低位，这表明“有必要重新关注自然语言处理这一基础性第一步的研究”。[26]

[1] https://huggingface.co/datasets/rcds/MultiLegalSBD [2] https://huggingface.co/models?search=rcds/distilbert-sbd 和 https://github.com/tobiasbrugger/MultiLegalSBD/tree/master/models [3] https://github.com/tobiasbrugger/MultiLegalSBD

2.1 法律领域中的句子边界检测

Savelka [29] 等人在英语语言中继续开展这项研究，整理了一个由美国裁决组成的法律数据集。在该数据集上测试现有系统时，他们报告的F1 分数介于75%到78%之间。将这些系统在该数据集上进行训练或适配后，其F1分数提高到了80多分，但仍低于它们在更标准领域中的相应性能 [26], ，这展示了英语法律领域中最先进句子边界检测系统的欠佳性能。为改善这一问题，他们训练了多个条件随机场模型，以及一个基于手工制定的规则的模型；手工制定的规则模型的F1 分数为79%，而条件随机场模型的F1 分数最高可达96%。此外，他们还制定了一套公开可用且全面的法律文本句子边界标注指南，我们以此作为制定自身指南的基础。

Sanchez [27] 在同一数据集上进行了实验，使用 Punkt 模型 [15] 得到的 F1 分数为 74%；针对该数据集进行适配后，性能略有提升。他们还训练并评估了条件随机场和神经网络（NN）模型，报告的 F1 分数分别高达 98.5% 和 98.4%。我们的多语言模型在同一数据集上的 F1 分数介于 95.1% 至 97% 之间。

同样，Glaser 等人 [10] 整理了一个德国法律数据集，其中包含法律和判决；我们的研究也采用了类似的数据分布。他们评估了现有句子边界检测系统的基线性能，并将其与基于上述数据集训练的条件随机场和神经网络模型进行了比较。他们的研究结果显示，现成系统的 F1 分数介于 70% 至 78% 之间，这支持了现有句子边界检测系统在法律数据上表现不佳这一观点。条件随机场和神经网络模型的 F1 分数最高可达 98.5%。然而，据报告，将这些模型应用于此前未见过的德国法律文本时，性能显著下降，分数低至 81.1%。我们的多语言模型在德国数据集上的 F1 分数介于 91.6% 至 97.6% 之间。

2.2 其他领域中的句子边界检测

在金融领域，Du 等人 [7] 尝试了结合条件随机场层的双向长短期记忆网络（双向长短期记忆网络）模型，以及基于 Transformer 的模型 BERT [6] ，并比较了它们的性能；他们将句子边界检测视为序列标注任务，以便从含噪金融文本中提取有用的句子。他们证明，BERT 在包括 F1 分数在内的所有评估指标上都显著优于双向长短期记忆网络-条件随机场模型。他们在研究中还强调了这样一个事实：“与自然语言处理领域中一些更受欢迎的子任务和主题相比，句子边界检测在过去几十年中受到的关注要少得多。”

Schweter 和 Ahmed [31] 比较了长短期记忆网络（长短期记忆网络）、双向长短期记忆网络和卷积神经网络（卷积神经网络）与 OpenNLP [4] 在 Europarl[16], SETimes [33] 和莱比锡语料库 [11] 上的句子边界检测任务中的性能；这些语料库包含大约 10 种不同语言，展示了其模型作为稳健且与语言无关的句子边界检测系统的应用。

2.3 法律领域的多语言数据集

Niklaus [25] 等人介绍LEXTREME，这是一个新颖的多语言基准测试数据集，包含涵盖24种语言的11个数据集，旨在评估自然语言处理模型在法律任务上的表现。作者评估了五种常见的多语言语言模型，为研究人员提供可用作比较依据的基准测试。Savelka 等人 [30] 研究了在序列标注模型中应用多语言句子嵌入，以促进跨语言、司法管辖区及其他法律领域的迁移。他们展示了在不同情境中复用已标注数据的积极成果，这有助于开发更具韧性且泛化能力更强的模型。此外，他们还使用这些模型创建了一个规模庞大的新标注法律文本数据集。Chalkidis 等人 [3] 推出了 MultiEURLEX，这是一个多语言、多标签法律文档分类数据集，包含 65000 部欧盟法律。Aumiller 等人 [1] 推出了 EurLexSum，这是一个根据 Eur-Lex 数据整理而成的多语言摘要数据集。Niklaus 等人 [21, 24]推出了 Swiss-Judgment-Prediction，这是一个来自瑞士联邦最高法院的多语言判决预测数据集。

3 数据集

我们为三个涵盖多种语言的法律数据集标注了句子跨度，分别为法语、意大利语和西班牙语数据集，每个数据集约包含 20,000 个句子，判决和法律各占一半。我们选择了多个法律领域，以涵盖广泛的内容。法律文本包括宪法、《民法典》的一部分和《刑法典》的一部分，其中宪法仅用于评估。判决包括来自不同法律领域和来源的法院判决。我们还标注了一个规模较小的葡萄牙语数据集，约含 1,800 个句子，并划分为与其他数据集相同的子集。该数据集用于零样本实验。

此外，我们对两个公开可用的数据集进行了标准化并将其整合进我们的数据集中，以进一步提高数据集的多样性：一个由美国裁决构成的英语法律文本集合 [29]，以及一个包含法律和判决的德语数据集 [10], 。

图 2展示了我们数据集中句子长度的分布，呈现了法律文本和判决中不同句子长度（以词元计）的相对频率，组距为 5。我们采用了激进的分词器，因此每个句子的词元数比通常情况更多。为清晰起见，我们没有纳入长度超过 101 个词元的句子，这些句子仅占全部句子的 ~2%（2634 个）。仅有 26 个句子的长度超过 512 个词元。

对于每种语言，我们都使用随机抽样将数据集划分为三个部分：训练集、测试集和验证集。测试集和验证集各占数据集的 20%。每个模型都在训练集上训练，我们报告其在测试集上的性能。数据集的部分统计数据和信息见表 1。

3.1 标注

人工标注员负责修正自动句子边界检测系统[5] [29] 预测的句子片段。该系统基于条件随机场，并使用 Savelka 等人依据标注指南进行标注的数据进行训练 [29]。这有助于提高我们标注的质量和一致性。此外，我们还采用了一套深受上述指南影响的实用规则集，以辅助标注员完成标注过程，降低任务复杂度，并帮助提供可靠且有充分依据的数据。该

[4]https://opennlp.apache.org/

[5]https://github.com/jsavelka/luima_句子边界检测

表1：各语言和子集的数据集统计 വിവര

语言	子集	句子	词元	文档数	来源
法语	判决	9971	342469	315	Niklaus 等人 [21]
	Laws	11055	334453	3	Wipolex
意大利语	判决	10129	340041	183 + 60	Niklaus 等人 [21] + Multi-Legal-Pile（MLP）
	Laws	10849	301466	3	Jus.unitn.it
西班牙语	判决	10656	356681	20 + 84	Wipolex + MLP
	Laws	11501	229240	3	Wipolex
葡萄牙语	判决	759	20590	6	Wipolex
	Laws	1010	25947	3	Wipolex
德语	判决	21409	506009	131	Glaser 等人 [10]
	Laws	20330	484816	13	Glaser 等人 [10]
英语	判决	25899	712433	80	Savelka等人 [29]
总计	法律与判决	133568	3654145	906	

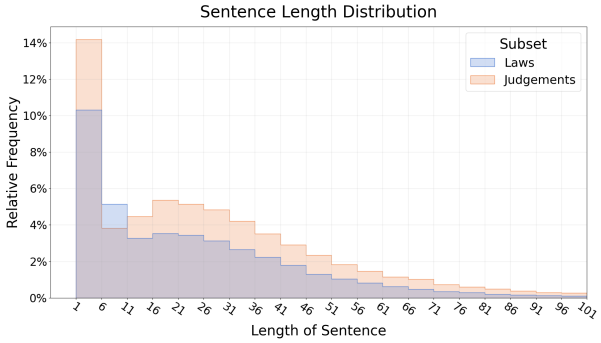

图2：词元中的句子长度分布 **ns**

规则集详见第 3.1.1 节，其中列出了最重要的句子结构，并附有示例。

这些文档使用 Prodigy 进行标注（https://prodi.gy/）。由于 Prodigy 要求文本预先进行分词，因此对输入文本应用了定制分词器，详见第3.2节。我们决定标注完整的句子范围，而不仅仅是序列中的第一个和最后一个词元，以鼓励标注人员阅读文本，而不是浏览文本寻找句末字符。为便于标注，法律被拆分为较小的片段，每个片段包含一至三条条款；而判决仅在超过 ~15000 个字符时才会拆分，因为 Prodigy 无法处理更长的文档。

3.1.1 法律句子结构。 本节简要介绍法律文本中最重要的句子结构，这些结构很大程度上受到Savelka等人研究的影响。 [29],随后给出一个法语示例。

标准句 主语、宾语和动词的语序正确，且序列中的最后一个词元是句末标点。• 他已成为一名独立工程师。**语言变换句** 与标准句相似，但会进行轻微变换，例如改变语序。

- 事实上，一切迹象都表明，她注定会陷入过度负债，继而破产。

标题 决定文本的结构，并展现文档各部分之间的关联，因此传达了有关文本整体结构的重要信息。

- 事实与法律方面的考虑
- 据此作出裁决
- 声明

数据字段 提供字段的名称和数据。此类内容被标注为一个句子，例如，英语中的“民事庭：马德里”与“民事庭位于马德里”含义相近。

- 上诉编号：*1231/2015*

括号在法律文本中经常出现，通常与引文连用。我们将括号标注为其所属的句子。括号内的内容不单独标注，如以下包含单个句子的示例所示：

- 由于其最后住所位于国外，实际上无法享有康复措施（《瑞士与南斯拉夫社会保障公约》，*1962*年*6*月*8*日，第*8a*条第*1*款）。

冒号不应标注为句末标点，除非冒号后紧接换行。理由是，以冒号结尾且后接换行的内容通常会引出列表或引文块，应与引导它的句子分别标注。

列表的标注方式因类型而异。对于列表项为不完整句子的列表（通常以分号结尾），整个列表标注为一个句子。以下示例由两个句子组成：冒号前的引导句，以及从“1°”到句号的内容。

- 在法院规定的履行公益劳动期限内，被判刑者必须遵守以下监管措施：*1°* 响应刑罚执行法官的传唤；*2°*（.....）患有可能危及其他劳动者的疾病。

不过，如果列表项本身是句子，则列表编号（或字母）和列表项都分别标注为一个句子，因为它们表达了不同的意思。下面的示例中有 3 个句子（引导语、列表编号、列表项）。

- 法律理由如下：
  1.- 在联邦审级中，仅可审查上诉委员会是否有正当理由要求上诉人预缴*500*瑞士法郎的费用，以支付一审上诉程序的费用。

省略号用于表示句子的一部分或文档的一部分被省略。以下示例展示了省略号的用法。第一个省略号单独标注，因为它表示有句子缺失。第二个省略号表示该句的一部分被省略，因此不单独标注。

- *(...)X.*的过错极其严重，因为归咎于他的行为（...）是在某种冷酷态度下实施的，令人无法容忍，并且对受害者而言难以忍受。

脚注 / 尾注向读者提供额外信息。尾注和脚注的标记（如数字或字母）应始终标注为位于句子范围内，即使它们出现在句末标点之后。如下例所示，以下序列只有一个句子，“(2)”是其标记：

- 法律仅对未来发生效力；不具有追溯力。(2)

此外，以编号列表形式出现的尾注应按照列表的标注指南处理。如下例所示，“(2)”构成一个句子，后面接着一个普通句子：

- (2) 议会议员的更替依照第25条的规定进行。

3.2 分词器

我们实现了一种基于正则表达式的激进分词器，用于将文本切分为词元；其他研究中也使用了这种分词器 [10, 29]。该分词器适用于所有语言。单词、数字和换行符、空白等特殊字符均被分隔为独立序列。这样做是为了确保在句子边界检测过程中不会丢失任何重要信息（例如，表示句子边界的换行符）。以下展示了一个示例；为清晰起见，省略了分词后的空白：

- *D._* *于* *1979* *年* *7* *月* *16* *日入学。* • *D* | . | _ | *est* | *entré* | *à* | *l* | ’ | *école* | *le* | *16* | *juillet* | *1979* | .

4 实验设置

我们开展了一系列实验，以回答第 1.2 节提出的研究问题。首先，我们比较了选定的现有模型，以确立基线性能。其次，我们基于条件随机场、双向长短期记忆网络-条件随机场和Transformer模型，训练并评估了各种单语言和多语言模型，并将其与基线进行比较。最后，我们在零样本实验中评估了多语言模型在未见数据上的性能。

4.1 基线系统

我们对多个广泛使用的系统进行了全面评估，这些系统采用了各种技术，包括 CoreNLP、NLTK、Stanza 和 spaCy，并将其作为我们的基线。下一节将详细描述每个系统。

4.1.1 NLTK。 由 Kiss 和 Strunk [15]创建的完全无监督句子边界检测系统。该系统的核心思路是，大多数错误预测的句子边界都源于缩写后的句号。因此，系统通过考察文本中每个词元的长度、搭配关系、内部句号，以及不带结尾句号的缩写出现情况，来识别缩写。我们测试了一个预训练模型，以及一个在我们的数据上训练的模型。

4.1.2 CoreNLP。 斯坦福 CoreNLP 工具包 [18], 中的一个基于规则的系统，根据句号、问号或感叹号等事件来预测句子边界。

4.1.3 Stanza。 一种基于双向长短期记忆网络模型的多语言系统[25]。我们仅使用其自然语言处理流水线的第一部分，即分词器。它将分词和句子切分联合处理，将其视为字符序列标注问题，预测某个字符是否为词元或句子的结束位置。

4.1.4 spaCy。 一种多语言系统 [12]，使用卷积神经网络和Transformer模型等技术的预训练模型。就我们的用途而言，仅使用了分词器和句子切分器。

4.2 我们的模型

基于第 2 节所介绍的工作，我们选择测试基于条件随机场、双向长短期记忆网络-条件随机场和Transformer模型的模型。我们将在以下小节中进一步介绍这些模型。为进行测试，我们训练了[6] 并评估了每种语言的单语模型，以及使用除葡萄牙语之外的所有语言训练的多语言模型：分别针对法律、判决，以及将这两种类型结合起来的模型。

4.2.1 条件随机场。 第 3.2 节中的分词器对输入文本进行了分词，包括空白字符。每个词元都被转换为表示该词元的简单特征列表，并加入了预定义窗口中该词元周围的词元特征。受 Glaser 等人启发，每种特征的窗口大小各不相同。 [10] 以及 Savelka等人 [29], , 如表 2所示。我们遵循 Lin 等人的方法，使用“BILOU”系统对输入数据进行标注。 [17]。

为训练我们的CRF模型，我们使用了 python-crfsuite[7] 实现。每个模型训练了 100 次迭代，正则化参数分别设为 C1 = 1 和 C2 = $1e^{-3}$，采用 L-BFGS 算法，并纳入所有可能的特征转移。

4.2.2 双向长短期记忆网络-条件随机场。双向长短期记忆网络将两个方向相反的长短期记忆网络连接到同一个输出，从而能够同时捕捉过去和未来状态的信息。每个长短期记忆网络的输出会拼接成每个输入词元的表示。对于双向长短期记忆网络-条件随机场模型，条件随机场层连接到双向长短期记忆网络的输出，并使用上述表示作为特征来预测最终标签。

[6]图形处理器：NVIDIA GeForce RTX 3060 TI，中央处理器：Intel Core i5-8600K CPU @ 3.60GHz [7]https://pypi.org/project/python-crfsuite/

表2：条件随机场特征的描述

特征	描述	窗口
特殊	每个词元都按照以下翻译进行分类：句子结束词元标记为“结束”，左括号和右括号标记为“左括号”分别为“右括号”和将换行字符标记为“Newline”，将缩写字符标记为“Abbr”，而其余部分标记为“No”。	10
小写	小写形式的词元。	7
长度	词元的长度。	7
签名	每个字符都通过以下翻译来表示：小写和大写字符分别改写为“c”和“C”，数字写作“N”，特殊字符标记为“S”。	5
较低	首个字符是否为小写字母case.	3
上层	首字符是否为大写case.	3
数字	词元是否为数字。	3

我们使用 Bi-LSTM-CRF[8] 库来训练模型。词嵌入维度设为 128，隐藏维度设为 256，最大序列长度为 512。批量大小为 16，学习率为 0.01，权重衰减为 0.0001。每个模型训练 8 个轮次，并保存验证集损失最小的模型。我们从文档中提取训练所用的词嵌入。为了标注训练数据，我们采用了第 4.2.1 节所述的“BILOU”标注系统。训练时，我们将人工标注的句子组合成批次，每批词元数限制为 512，以模拟较长的段落。

4.2.3 Transformer。 Transformer 是一种利用自注意力机制的神经网络（NN），用于在进行预测时衡量输入不同部分的重要性。BERT 等 Transformer 模型使用多层编码器 [34]，通过在所有层中同时以左右上下文为条件，对深度双向表示进行预训练 [6]。因此，我们可以通过添加一个额外的输出层，将 Transformer 模型微调到句子边界检测任务上。在本研究中，我们针对句子边界检测任务中的所有语言，使用了一个基于 DistilBERT [28],的预训练模型[9]；DistilBERT 是 BERT 的一个更小、更轻量的版本。[10] 我们使用 PyTorch[11] 和 Accelerate[12]，配合 Adam 优化器，对模型训练了 5 个 epoch，批量大小为 8，学习率为 $2e^{-5}$。

DistilBERT 的一个局限是输入长度上限为 512 个词元，因为自注意力机制的运行时间会随着序列长度呈平方级增长。由于 DistilBERT 依赖 WordPiece 分词器 [32],将文本拆分为子词，因此每个序列中的词元数量会更多，这使得这一问题更加突出。因此，为了绕过 512 个词元的限制，我们使用金标准标注将每个文档拆分为句子。然后，将连续的句子依次添加到一个集合中，直到总长度尽可能接近词元上限。接下来，模型预测每个集合中的句子边界。超过 512 个词元的句子会被截断。[13] 这一方案显而易见的缺点是，输入文本必须事先拆分为句子或较短的片段，因此很难将 BERT 模型应用于未知文本。

未来的工作可以探究串联句子边界检测模型是否可行（即先在输入文本上应用条件随机场模型，将文本切分为长度小于 512 个词元的片段，再应用基于 Transformer 的模型）。另一种方案是使用经过预训练、支持更长输入文本的 Transformer 模型，并利用注意力机制，使计算量随序列长度线性增长，例如 Longformer [2].[14]

4.3 评估

句子边界检测任务的一个特点是，非句子边界标签天然占据多数，因为每个句子最多只有两个句子边界。因此，为了更准确地评估我们的模型，我们采用了常用的评估指标——精确率（P）、召回率（R）和 F1 分数（F1）。尽管句子边界检测任务在专业领域中尚未得到解决，但与问答或摘要等其他自然语言处理任务相比，它相对容易。由于句子边界检测是一项预处理任务，因此必须取得更高的分数，才能防止错误传播到下游任务。由此，我们认为，最先进的句子边界检测模型的 F1 分数应达到较高的九十分位数，才能在实践中发挥作用。

在评估过程中，我们让模型预测每份文档中的句子跨度。这些标注出的跨度由我们的分词器进行分词（第 3.2 节）。随后，根据每个词元是否为句子边界，为其赋予一个二元值。这将预测的句子跨度或边界与所用的分词器解耦，因为某些模型的分词器可能会将略有不同的词元指定为句子的第一个或最后一个词元，以下法语示例将对此作进一步说明：*“C’est en outre ...”*。我们的分词器会将“C”指定为序列中的第一个词元，而其他分词器可能会将“C’”甚至“C’est”指定为第一个词元。与金标准标注相比，这会导致句子边界被错误预测，尽管预测实际上是正确的。

使用 Scikit-Learn 比较每种文档类型的真实标签和预测标签，以计算二元F1分数。各子集的分数取平均值：“法律”包括刑法典、民法典和宪法；“判决”则包括各种法院判决。

我们对每个条件随机场模型训练一次，并使用随机种子对双向长短期记忆网络-条件随机场模型和 Transformer 模型各训练 5 次，报告平均性能及标准差。除非另有说明，报告的数值均为二元F1分数。

[8]https://github.com/jidasheng/bi-lstm-crf [9] https://huggingface.co/distilbert-base-multilingual-cased [10]为提高效率，我们使用了一个更小的模型；未来的工作建议使用更大的模型。[11] https://pytorch.org/ [12]https://huggingface.co/docs/accelerate/index

[13]这导致了一些句子边界预测错误，不过这种情况只发生了几次，因此对总体得分影响不大。[14]遗憾的是，据我们所知，目前还不存在经过多语言预训练的高效 Transformer 模型。

表3：基线模型和多语言模型在所有语言以及葡萄牙语零样本实验中的平均（±标准差）F1分数。最佳分数以粗体显示。

语言	法语		西班牙语		意大利语		英语	德语		葡萄牙语（零样本）	
Type 模型	判决	Laws	判决	Laws	判决	Laws	判决	判决	Laws	判决	Laws
CoreNLP	74.7	76.7	71.4	89.0	79.8	75.6	81.7	69.0	64.0	-	-
NLTK	72.5	75.8	70.2	89.2	72.3	66.3	77.2	72.3	73.8	64.9	57.0
NLTK-train	82.9	75.8	72.1	81.6	84.8	77.5	84.9	74.2	73.5	71.7	64.3
spaCy	86.6	67.2	60.0	70.3	73.9	73.7	79.7	87.5	67.0	59.0	77.7
Stanza	81.9	81.0	83.2	90.2	85.7	87.4	92.3	72.6	64.7	88.6	73.4
CRF	97.8	98.1	94.8	98.9	97.3	97.7	95.1	95.2	91.6	90.2	78.6
双向长短期记忆网络-条件随机场	97.6±0.3	**98.5**±0.2	97.3±0.1	**99.3**±0.2	97.8±0.1	**99.2**±0.1	95.4±0.3	**97.2**±0.2	97.5±0.5	93.0±0.6	73.2±3.3
Transformer	**98.3**±0.1	98.1±0.2	**97.8**±0.1	99.0±0.0	**98.3**±0.1	99.1±0.1	**97.0**±0.1	92.9±0.2	**97.6**±0.1	**93.6**±0.3	**91.3**±1.1

5 结果

5.1 基线模型

第 4 节中，我们的数据集各语言的基线模型性能汇总于表 3上半部分。

基线模型的结果明显低于 Read 等人报告的用户生成内容得分，这支持了这样一个假设：对于所有测试的语言，开箱即用模型在法律数据上的表现都不尽如人意。表现差异的一部分可能是由第 3.1 节所述的特殊句子结构造成的，另一部分则归因于法律文本本身的挑战性。

值得关注的是，在大多数语言中，NLTK 与 NLTK-train 之间存在差距，因为训练 NLTK 后，其识别并正确预测缩写的能力有所提升。这表明，缩写是法律文本具有挑战性的一个方面。需要注意的是，与其他模型相比，spaCy 对句子的定义略有不同：通常，当两个句子由换行符分隔时，换行符不会属于任何句子跨度；而 spaCy 会将其包含在第二个句子的跨度中。这会导致错误预测，尽管 spaCy 正确识别出其中有两个句子。因此，spaCy 获得的分数低于预期。

5.2 单语模型

我们在表 4 中报告了所训练单语模型的性能。每个模型都在同一种语言上进行训练和测试。

我们观察到，将每个模型应用于其训练子集时，其性能在几乎所有语言上都达到了九十多分，显著优于第 5.1 节中的基线模型，并且与已报告的英语新闻条款上的句子边界检测系统性能相当 [26]。我们的模型在英语上的表现也与已报告的条件随机场和卷积神经网络的性能相近 [27, 29], ，在德语数据集上也与条件随机场和神经网络的性能相近 [10]。

比较在一个子集上训练、并在另一个未见过的数据集上评估的模型性能，即进行零样本实验，可以发现，Transformer模型在大多数语言上优于条件随机场和双向长短期记忆网络-条件随机场；当在判决上训练、在法律上评估时，其在意大利语数据集上的性能降至81.8%，但仍可与最佳基线模型相媲美。不出所料，模型在零样本实验中的性能几乎总是低于其在训练所用子集上的性能。这一差距可以由判决与法律之间在写作和格式风格上的巨大差异来解释，其中Transformer模型最擅长在两个子集之间泛化知识。我们进一步推测，与在法律上训练相比，模型在判决上训练时更容易将其知识泛化到不同领域，从而在未见数据上取得更高的分数。这里的一个因素可能是，判决中的法律文本包含更多样的句子结构，而法律通常会重复使用相同的结构。

在西班牙语数据集上，以法律为训练数据、以判决为评估数据时，条件随机场和双向长短期记忆网络-条件随机场模型的性能尤其差，分数分别低至 43.4% 和 54.3%。我们推测，与 Transformer 模型相比，这两种模型对不同领域的泛化能力较弱。

总而言之，尽管同时使用法律和判决进行训练并不总能带来绝对最佳性能，但这种做法最为稳健，也不会导致性能下降。

5.3 多语言模型

我们在法律和判决数据上训练的多语言模型的性能报告于表 3的下半部分。每个多语言模型均使用除葡萄牙语之外的所有语言进行训练。

多语言模型明显优于基线模型，F1 分数最高可达 99.2%。双向长短期记忆网络-条件随机场和 Transformer 模型的表现都非常出色，其中 Transformer 模型在判决数据上的表现略好，而双向长短期记忆网络-条件随机场模型在法律数据上的表现略好。条件随机场模型也紧随另外两种模型之后，大多取得了 90 多分的高分。将多语言模型与单语模型的性能进行比较可以发现，使用规模大得多的数据集进行训练并不会造成性能损失；多语言模型的表现与各自语言的单语模型相当，而 Transformer 和双向长短期记忆网络-条件随机场模型甚至在每种语言上的表现都优于单语模型。

5.4 葡萄牙语数据上的零样本实验

我们进行了一个更具挑战性的实验，在葡萄牙语数据上评估多语言模型，并将其与基线进行比较。图 1概述了整体情况，而表 3 详细列出了判决和法律与基线之间的差异。

表4：单语模型在各自语言上的平均 F1分数（±标准差）。最佳分数以粗体显示。

语言		法语		西班牙语		意大利语		英语	德语	
模型	Type 在……上训练	判决	Laws	判决	Laws	判决	Laws	判决	判决	Laws
CRF	判决	97.9	73.2	97.0	98.3	**98.5**	95.6	96.8	97.8	76.5
	Laws	78.5	**98.8**	54.3	**99.6**	88.6	**99.6**	-	75.8	**97.7**
	法律 + 判决	97.8	**98.8**	97.0	99.5	98.3	99.5	-	97.2	97.2
双向长短期记忆网络-条件随机场	判决	97.3±0.3	56.7±3.0	94.7±0.5	92.1±0.9	95.9±0.3	71.3±2.2	**97.3**±0.4	97.0±0.3	76.9±0.4
	Laws	66.1±4.2	97.9±0.2	43.4±6.8	98.7±0.2	74.1±1.2	98.4±0.2	-	71.9±2.5	97.3±0.3
	法律 + 判决	97.0±0.4	98.1±0.1	95.6±0.5	98.9±0.4	96.2±0.2	98.2±0.1	-	97.2±0.2	97.6±0.2
Transformer	判决	98.2±0.1	84.7±1.2	96.9±0.2	96.9±0.4	97.8±0.2	81.8±0.9	96.5±0.1	98.0±0.2	87.2±0.4
	Laws	92.4±0.5	97.6±0.4	89.5±0.6	97.1±3.7	89.4±0.7	98.8±0.5	-	89.4±0.5	97.4±0.1
	法律 + 判决	**98.4**±0.1	98.2±0.2	**97.3**±0.1	99.0±0.1	97.1±0.3	99.1±0.1	-	**98.3**±0.1	97.5±0.2

表 5：单语和多语言模型在未见过的葡萄牙语数据上的平均 F1 分数

模型	CRF		双向长短期记忆网络-条件随机场		Transformer	
Type 模型语言	判决	Laws	判决	Laws	判决	Laws
法语	79.3	75.4	25.5	51.7	82.5	87.1
西班牙语	**91.5**	79.4	80.3	**73.5**	88.0	**94.0**
意大利语	81.8	**83.3**	12.6	64.8	70.0	73.7
英语	90.6	72.1	80.6	62.4	87.6	89.9
德语	59.0	25.2	43.6	30.3	79.9	71.1
多语言	90.2	78.6	**93.0**	73.2	**93.6**	91.3

对于判决，性能良好，F1 分数介于 90.2% 和 93.6% 之间，与用户生成内容相当 [26], 且优于大多数基线。然而，对于法律文本，只有 Transformer 模型的得分达到 90% 出头，而条件随机场和双向长短期记忆网络-条件随机场分别降至 78.6% 和 73.2%，与我们通常的基线值相近。Transformer模型的大规模多语言预训练可能使其对分布偏移更具鲁棒性，从而比条件随机场或双向长短期记忆网络-条件随机场更好地迁移到未见过的语言。

葡萄牙语法律文本的写作和格式风格较为复杂，这或许可以解释法律与判决之间的差异，这种差异体现在葡萄牙语基线性能低于通常水平。双向长短期记忆网络-条件随机场的性能下降，也可能是由于训练中缺少葡萄牙语词嵌入所致，因为我们仅从训练数据中提取了嵌入。为提升双向长短期记忆网络-条件随机场模型的性能，未来研究可以探索在训练过程中添加葡萄牙语词嵌入，或使用规模更大的多语言嵌入词汇表。为提升 Transformer 模型的性能，可以考虑在句子边界检测任务上微调 XLM-RoBERTa [5] 等较大型的预训练模型，因为与 mBERT [6] 或 DistilBERT [28] 模型相比，这些模型的跨语言迁移能力显著更强。

在评估使用完整单语数据集训练的单语和多语言模型对先前未见过的葡萄牙语数据（表 5）的有效性时，我们观察到，多语言模型的性能优于相应的单语模型

在大多数语言中，西班牙语是一个显著的例外。我们推测，性能差异源于西班牙语和葡萄牙语之间密切的语言联系，这使西班牙语单语模型能够在跨语言迁移中表现出色。然而，对于其他在语言上与西班牙语不太接近的语言，预计多语言模型的表现将优于单语模型。

5.5 推理时间

表6 报告了我们在法律和判决上训练的多语言模型的推理时间。我们分别在图形处理器（NVIDIA GeForce RTX 3060 TI）和中央处理器（Intel Core i5-8600K CPU @ 3.60GHz）上测量了三次推理时间，并报告其平均值。由于没有显著的离群值，因此未报告标准差。值得注意的是，Transformer模型在图形处理器上的推理时间显著缩短。然而，条件随机场的评估并未从图形处理器中受益，因为它采用顺序运算。

表6：各多语言模型预测包含 ~130000 个句子的整个数据集以及一个句子所需的平均推理时间，单位为分钟（min）、秒（s）和毫秒（ms），测量平台为图形处理器和中央处理器

	完整数据集（~130000 个句子）		一个句子	
模型	CPU	GPU	CPU	GPU
CRF	11分钟57秒	-	~5.37 毫秒	-
双向长短期记忆网络-条件随机场	10 分 6 秒	9 分 23 秒	~4.54 毫秒	~4.21 毫秒
Transformer	34 分 26 秒	9 分 18 秒	~15.47 毫秒	~4.18 毫秒

综合第5.2、5.3和5.4节所呈现的结果，考虑到推理时间和易用性，只要有可用的图形处理器用于推理，在大多数情况下，我们都可以推荐使用多语言 Transformer 模型。对于特定语言的任务或需要处理较长输入文本的任务，我们建议使用相应语言的条件随机场模型，尽管与双向长短期记忆网络-条件随机场模型和 Transformer 模型相比，这些模型的设置时间更长。

5.6 错误分析

我们检查了随机样本——多语言模型预测的葡萄牙语数据集的三分之二（8 份判决、20 部法律）。

用于葡萄牙语文本零样本实验的Transformer模型。我们按照第5.5节中的建议选择了多语言Transformer，并选择了葡萄牙语数据集，因为该模型在其他数据集上的表现已经非常出色。

标准句子边界很少被漏掉，模型在这方面表现尚可；不过，我们发现了一些常见错误的来源。我们讨论了一些示例，其中 |T| 和 |P| 分别表示真实句子边界和预测句子边界。如下例所示，许多错误源于引文和括号：

- (Bittar, Carlos Alberto. |P| 著作权。 |P| 里约热内卢：Forense Universitária，2001年，第143页） |T| |P|

在此示例中，我们有一个引文句子，其中的句号被错误地预测为引文内部的句子边界。

错误的另一个来源是数据字段和标题，因为它们通常缺乏足够的提示（例如句末标点），使模型难以将其识别为此类内容：

(1) 主审法官：SIDNEI BENETI 部长 |T|(2) 判决书 |T|

模型未能在两个序列的末尾预测出句子边界。上述示例所展示的错误主要源于我们特别定义的句子结构（第 3.1.1节），以及法律句子边界检测任务本身的挑战性。

另一类错误由葡萄牙语中不同的格式风格和模型不熟悉的用词造成，例如：

(1) 合议庭一致同意，根据主审法官 |P| 法官的意见，准许特别上诉。|T| |P| (2) 尊敬的 |P| 上诉法院法官 MAURÍCIO PESSOA（主席）、CLAUDIO GODOY E GRAVA BRAZIL。 |T| |P|

在（1）中，出现了缩写“Sr(a)”，模型未能将其识别为缩写，因此将句点标记为句子边界。（2）中也出现了类似错误，涉及缩写“Exmos”。

5.7 局限性

由于标注人员的语言能力所限，我们只对两个语言群组（日耳曼语族和意大利语族）的数据进行了标注。因此，我们所涵盖的语言在词汇上高度重叠，使得跨语言迁移相对容易。未来的研究可以考察更多不同语言群组的法律文本，以构建对语言分布偏移更具鲁棒性的系统。

标注人员以德语为母语，并具备中级法语水平。由于意大利语、西班牙语和葡萄牙语与法语相似，而且句子边界检测任务在很大程度上具有结构性，因此能够完成标注。不过，由相应语言的母语者执行标注，或许能进一步提高标注质量。另一方面，由一名标注人员（如本研究所采用的做法）标注整个数据集，则有助于提高不同语言间的一致性。

由于经费有限，我们只安排了一名标注人员进行标注。由第二名标注人员对标注结果进行验证，或许能进一步提高标注质量。

<escape>对第 5.6 节中的定性错误分析进行定量补充，可能为进一步改进系统提供更具体且可操作的证据。为实现这一目标，需要更加详细的</escape>

对句子类型进行标注会很有帮助，这样便可以对句子进行统计，从而获得表现最差的句子类型的定量结果。

6 结论与未来工作

6.1 研究问题的答案

研究问题1： 现有句子边界检测系统在法语、西班牙语、意大利语、英语和德语法律数据上的性能如何？现有句子边界检测系统在所有测试语言中的表现都不理想，低于 Read 等人报告的分数 [26]，这些分数基于用户生成的 d

内容，这表明句子边界检测在法律领域尚未得到解决。

研究问题2： 通过基于条件随机场、双向长短期记忆网络-条件随机场和 *Transformer* 的模型进行训练，我们在多大程度上能够提升这一性能？单语模型在所有测试语言上均取得了九十多分的最先进 F1 分数，与新闻文章上的报告分数相当 [26]。多语言模型的表现与单语模型相近，展现了使用更大的数据集进行训练的潜力。与条件随机场模型和双向长短期记忆网络-条件随机场模型相比，Transformer 模型表现出更强的跨领域迁移能力。**研究问题3：** 多语言模型在未见过的葡萄牙语法律文本上的性能如何，即在零样本实验中的表现如何？Transformer 模型在判决和法律子集上的表现尚可，F1 分数达到九十出头，展现出最强的跨语言迁移能力；条件随机场模型和双向长短期记忆网络-条件随机场模型在判决上的表现也不错，约为 90%，但在法律子集上的表现则降至基线值，这很可能意味着还需要进一步优化。

6.2 结论

在本研究中，我们整理并公开发布了一个多样化的法律数据集，其中包含6种语言的超过13万条经过标注的句子，为法律领域的进一步研究提供了支持。利用该数据集，我们表明，现有的句子边界检测方法在多语言法律数据上的表现较差，F1 分数最高也仅达到九十出头。我们训练并评估了单语言和多语言条件随机场、双向长短期记忆网络-条件随机场以及Transformer模型，在我们的数据集上取得了九十多分的二元F1分数，展现出当前最先进的性能。作为更具挑战性的任务，我们在零样本实验中使用多语言模型对未见过的葡萄牙语数据进行了测试，其中Transformer模型取得了九十出头的分数，大幅优于使用葡萄牙语文本训练的基线模型以及条件随机场和双向长短期记忆网络-条件随机场模型。我们公开发布了这些模型及其代码，供社区进一步使用和研究。

6.3 未来工作

通过进一步预处理输入文本，或许可以提升所有模型的表现，例如将换行符替换为空格，将特殊字符替换为更常用的等效字符，例如将双引号（“）替换为单引号（"). 此外，针对特定数据集进行全面的超参数优化，可能有助于提升多语言条件随机场和双向长短期记忆网络-条件随机场模型的表现。最后，Transformer模型或许能从面向法律领域的模型、更大型的预训练模型（如 BERT），或专为跨语言迁移任务设计的模型（如 XLM-RoBERTa）中获益。

使用多种语言的法律文本以及来自不同来源（如隐私政策和服务条款）的文档扩充数据集，或许能提升多语言模型的性能，尤其是在零样本场景下。如果在训练期间将句子跨度标注为其句子结构类型（例如"引文"（第3.1.1节）），而不是统一赋予单一标签，或许能观察到模型性能受到有趣的影响。

探究其研究中观察到的积极跨语言迁移是否也适用于不同语系的语言，例如匈牙利语。这一假设基于所研究语言的共同起源，正如第 5 节所述。

参考文献

[1] Dennis Aumiller、Ashish Chouhan 和 Michael Gertz。2022。EUR-Lex-Sum：法律领域长篇摘要的多语言与跨语言数据集。http://arxiv.org/abs/2210.13448 arXiv:2210.13448 [cs]。[2] Iz Beltagy、Matthew E. Peters 和 Arman Cohan。2020。Longformer：长文档 Transformer。arXiv: 2004 05150 [ ] 。cs Ilias Chalkidis、Manos [3] Fergadiotis 和 Ion Androutsopoulos。2021。MultiEURLEX——用于零样本跨语言迁移的多语言、多标签法律文档分类数据集。arXiv:2109.*00904 [cs]*（2021年9月）。http://arxiv.org/abs/2109.00904 arXiv：2109.00904。[4] Ilias Chalkidis、Manos Fergadiotis、Prodromos Malakasiotis、Nikolaos Aletras 和 Ion Androutsopoulos。2020。LEGAL-BERT：刚走出法学院的 Muppets。arXiv:2010.*02559 [cs]*（2020年10月）。http://arxiv.org/abs/2010.02559 arXiv：2010.02559。[5] Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。2020。大规模无监督跨语言表征学习。https://doi.org/ 10 48550 1911 02116 1911 02116 [ ] 。/ [6] 2019 arXiv。。arXiv：。cs Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。。BERT：用于语言理解的深度双向 Transformer 预训练。载于《2019年计算语言学协会北美分会会议论文集：人类语言技术》，第1卷（长篇与短篇论文）。计算语言学协会，明尼阿波利斯，明尼苏达州，4171 4186 10 18653 1 19 1423 –。https://doi.org/。/v /N - Jinhua [7] 2019 Du、Yan Huang 和 Karo Moilanen。。AIG Investments.AI 在 FinSBD任务中的表现：通过序列标注和 BERT微调进行句子边界检测。载于《第一届金融科技与自然语言处理研讨会论文集》。-，澳门，中国，81–87。[8] Dan Gillick。2009。句子边界检测与 U.S. 问题。载于《人类语言技术：2009年计算语言学协会北美分会年会论文集》，配套卷：短篇论文。计算语言学协会，博尔德，科罗拉多州，241–244。[9] Kevin Gimpel、Nathan Schneider、Brendan O'Connor、Dipanjan Das、Daniel Mills、Jacob Eisenstein、Michael Heilman、Dani Yogatama、Jeffrey Flanigan 和 Noah A. Smith。2011。Twitter 的词性标注：标注、特征与实验。载于《计算语言学协会第49届年会论文集：人类语言技术》。计算语言学协会，波特兰，俄勒冈州，美国，42–47。[10] Ingo Glaser、Sebastian Moser 和 Florian Matthes。2021。德国法律文件中的句子边界检测：。载于《第13届智能体与人工智能国际会议论文集》。SCITEPRESS - 科学与技术出版物，在线直播，— 选择国家 —，812–821。https://doi.org/ 10 5220 0010246308120821 。/ Dirk Goldhahn、[11] Thomas Eckart 和 Uwe Quasthoff。2012。在莱比锡语料库集构建大型单语词典：从100种语言到200种语言。《第八届语言资源与评测国际会议论文集》（LREC'12）-，-（2012），–。[12] Matthew Honnibal、Ines Montani、Sofie Van Landeghem 和 Adriane Boyd。2020。spaCy：Python 中的工业级自然语言处理。- -，-（2020），–。https://doi.org/ 10 5281 。Wenyue Hua、Yuchen Zhang、Zhe Chen、Josie Li [13] 和 Melanie Weber。2022。LegalRelectra：面向长距离法律文本理解的混合领域语言建模。https://doi.org/ 10 48550 2212 08204 2212 08204 。/arXiv。。[ ] arXiv：。cs。[14] Daniel Martin Katz、Dirk Hartung、Lauritz Gerlach、Abhik Jana 和 Michael James Bommarito。2023。法律领域的自然语言处理。[15] Tibor Kiss 和 Jan Strunk。2006。无监督多语言句子边界检测。《计算语言学》32（4）（2006年12月），485–525。https://doi.org/ 10.1162/coli.2006.32.4.485

[16] Philipp Koehn。2005。Europarl：用于统计机器翻译的平行语料库。载于《机器翻译峰会 X论文集：论文》，第-卷。-，普吉岛，泰国，79–86。[17] Chun-Wei Lin、Yinan Shao、Ji Zhang 和 Unil Yun。2020。基于潜变量条件随机场的增强序列标注。《Neurocomputing》403，-（2020年5月），–。https://doi.org/10.1016/j.neucom.2020.04.102[18] Christopher Manning、Mihai Surdeanu、John Bauer、Jenny Finkel、Steven Bethard 和 David McClosky。2014。斯坦福 CoreNLP 自然语言处理工具包。载于《计算语言学协会第52届年会论文集：系统演示》。计算语言学协会，巴尔的摩，马里兰州，55 60 10 3115 1 14 5010 –。https://doi.org/。/ [19] 2019 v /P - Ditty Mathew 和 Chinnappa Guggilla。。FinSBD 共享任务中的 AI Blues：金融领域 PDF 噪声文本的基于条件随机场的句子边界检测。载于《第一届金融科技与自然语言处理研讨会论文集》。-，澳门，中国，130–136。[20] Denis Newman-Griffis、Chaitanya Shivade、Eric Fosler-Lussier 和 Albert Lai。2016。临床领域句子边界检测的定量与定性评估。《AMIA 转化科学联合峰会论文集》。《AMIA 转化科学峰会2016》（2016年7月），88–97。[21] Joel Niklaus、Ilias Chalkidis 和 Matthias Stürmer。2021。Swiss-Judgment-Prediction：多语言法律判决预测基准测试。载于《2021年自然法律语言处理研讨会论文集》。计算语言学协会，蓬塔卡纳，多米尼加共和国，19–35。https: //aclanthology.org/ 2021 1 3 .nllp-。[22] Joel Niklaus 和 Daniele Giofré。2022。BudgetLongformer：我们能否以低成本从头预训练一个先进的法律语言模型？ https://doi.org/10.48550/arXiv.2211.17135 arXiv:2211.17135 [cs].[23] Joel Niklaus、Veton Matoshi、Pooja Rani、Andrea Galassi、Matthias Stürmer 和 Ilias Chalkidis。2023。LEXTREME：面向法律领域的多语言、多任务基准测试。https://doi.org/ 10 48550 2301 13126 2301 13126 。/arXiv。。[ ] arXiv：。cs。[24] Joel Niklaus、Matthias Stürmer 和 Ilias Chalkidis。2022。关于法律判决预测中跨 X 迁移的实证研究。载于《计算语言学协会亚太分会第二届会议暨第12届自然语言处理国际联合会议论文集》（第1卷：长篇论文）。计算语言学协会，仅在线， 32 46 2022 3 –。[25] https://aclanthology.org/。aacl-main。Peng Qi、Yuhao Zhang、Yuhui Zhang、Jason Bolton 和 Christopher D. Manning。2020。Stanza：适用于多种人类语言的 Python 自然语言处理工具包。载于《计算语言学协会第58届年会论文集：系统演示》。计算语言学协会，在线，101 108 10 18653 1 2020 14 –。[26] https://doi.org/。/v / .acl-demos。Jonathon Read、Rebecca Dridan、Stephan Oepen 和 Lars Jrgen Solberg。2012。句子边界检测：一个早已解决的问题？ 载于《COLING 2012论文集：海报》。-，孟买，印度，985–994。[27] George Sanchez。2019。法律文本中的句子边界检测。载于《2019年自然法律语言处理研讨会论文集》。计算语言学协会，明尼阿波利斯，明尼苏达州，31–38。https: //doi.org/ 10 18653 1 19 2204 。/v /W - Victor Sanh、Lysandre Debut、Julien [28] Chaumond 和 Thomas Wolf。2019。Distil-BERT，BERT 的蒸馏版本：更小、更快、更便宜、更轻量。ArXiv abs/1910.01108，-（2019），–。[29] Jaromir Savelka、Vern R Walker、Matthias Grabmair 和 Kevin D Ashley。2017。美国裁判决定中的句子边界检测。《TAL》58，21（2017），–。[30] Jaromir Savelka、Hannes Westermann、Karim Benyekhlef、Charlotte S. Alexander、Jayla C. Grant、David Restrepo Amariles、Rajaa El Hamdani、Sébastien Meeùs、Aurore Troussel、Michał Araszkiewicz、Kevin D. Ashley、Alexandra Ashley、Karl Branting、Mattia Falduti、Matthias Grabmair、Jakub Harašta、Tereza Novotná、Elizabeth Tippett 和 Shiwanni Johnson。2021。Lex Rosetta：跨语言、司法管辖区和法律领域迁移预测模型。载于《第十八届人工智能与法律国际会议论文集》。ACM，圣保罗，巴西， 129 138 10 1145 3462757 3466149 –。https://doi.org/。/ 。Stefan Schweter 和 Sajawel Ahmed。。Deep-EOS：通用神经网络句子边界检测。载于《第15届自然语言处理会议（KONVENS）论文集》，第-卷。-，-，5。[32] Xinying Song、Alex Salcianu、Yang Song、Dave Dopson 和 Denny Zhou。2020。快速 WordPiece 分词。- -，-（2020），–。https://doi.org/10.48550/ARXIV。2012.15524[33] Jorg Tiedemann。2012。OPUS 中的平行数据、工具与接口。载于《第八届语言资源与评测国际会议论文集》（LREC 2012）-，-（2012），–。[34] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017。Attention Is All You Need。https://doi.org/10.48550/arXiv.1706.03762 arXiv:1706.03762 [cs]

收稿日期：2023年2月16日；录用日期：2023年4月12日；修订日期：2023年5月2日