结合语义表示与上下文词表示识别越南语中的文本蕴含

Quoc-Loc Duong[1, 2], Duc-Vu Nguyen[1, 2], Ngan Luu-Thuy Nguyen[*],
[1,2]信息技术大学，越南胡志明市[2]越南国家大学，越南胡志明市
18521006@gm.uit.edu.vn {vund,ngannlt}@uit.edu.vn

arXiv:2301.00422v1 [cs.CL] 1 Jan 2023

摘要—**文本蕴含识别是一个重要问题，也是一个相当活跃的研究领域。针对这一问题，已有研究提出的方法十分多样，涉及许多不同方向。对于越南语而言，文本蕴含识别问题相对较新，但该问题在自然语言理解系统中发挥着至关重要的作用。目前，基于上下文词表示学习模型解决这一问题的方法已经取得了出色的结果。然而，越南语是一种语义丰富的语言。因此，本文希望展示一项实验：将通过SRL任务获得的语义词表示，与BERT相关模型的上下文表示相结合，用于解决文本蕴含识别问题。实验结果得出了关于语义表示对越南语理解的影响及其作用的结论。实验结果表明，具备语义感知的上下文表示模型，其性能比未纳入语义表示的模型高出约1%。此外，在越南语数据领域中的影响也高于英语中的影响。这一结果还表明，SRL对越南语文本蕴含识别问题具有积极影响。**

索引词——**文本蕴含识别、自然语言理解、语义角色标注、SemBERT、自然语言处理**

I. 引言

根据RTE-1会议，文本蕴含识别（RTE）任务定义为：给定两个文本片段，判断一个文本的含义是否可以从另一个文本中推断出来（即被另一个文本蕴含） [1]。这一问题在自然语言理解（NLU）领域中发挥着至关重要的作用。越南语具有多种含义，因此口语或书面文本中会产生歧义。然而，在实践中，这种歧义很少受到关注，因为人类能够很好地处理这一现象。歧义是指思想含混、模糊了事物之间界限的现象。歧义主要存在于越南语中，并且作为一种普遍现象不可避免
规则。许多自然语言理解问题，例如问答、文本摘要或信息抽取，在遇到歧义时都会面临处理困难。因此，文本蕴含识别有助于改善面对这一现象时的推理问题。此外，它还被用作评估自然语言处理系统整体语言理解能力的指标。

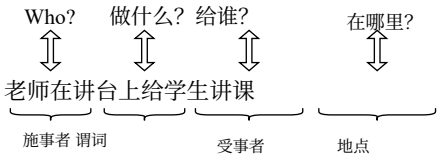

图 1. 越南语句子的语义角色标注。

另一方面，如前所述，越南语是一种语义丰富的语言，可能会产生歧义。因此，我们决定尝试通过语义角色标注（SRL）任务获取语义表示，并将其与基于Transformer 编码器模型获取的上下文表示相结合。这项实验可视为越南语文本蕴含识别问题的最新方法。另一方面，利用上下文表示是越南语文本蕴含识别问题的主要方法，并取得了出色的结果。

在某些语言学理论中，语义角色（SR）指名词、名词短语或句子其他成分在动作或状态中所承担的语义角色；这些成分统称为论元，而动作或状态由动词描述，也称为谓词。语义角色标注（SRL）旨在确定句子成分所承担的语义角色，以帮助厘清谓词与构成句子的论元之间的结构和关系。从本质上说，SRL是一个转换问题，但它在自然语言处理系统中发挥着关键作用。为了高效利用文本中编码的信息，系统必须能够检测所描述的事件。因此，语义角色有助于表示句子表层结构中的一般语义，从而支持简单推理，而这些推理无法通过解析树或其他方法得出。尤其是，SRL通过回答谁对谁做了什么、何时或为何等问题，帮助自然语言理解问题基于表层结构提升对自然语言的理解。另一方面，如前所述，越南语是一种语义丰富的语言；获取通用的语义表示可以提升自然语言理解问题（即文本蕴含识别）的模型性能。

[*]通讯作者

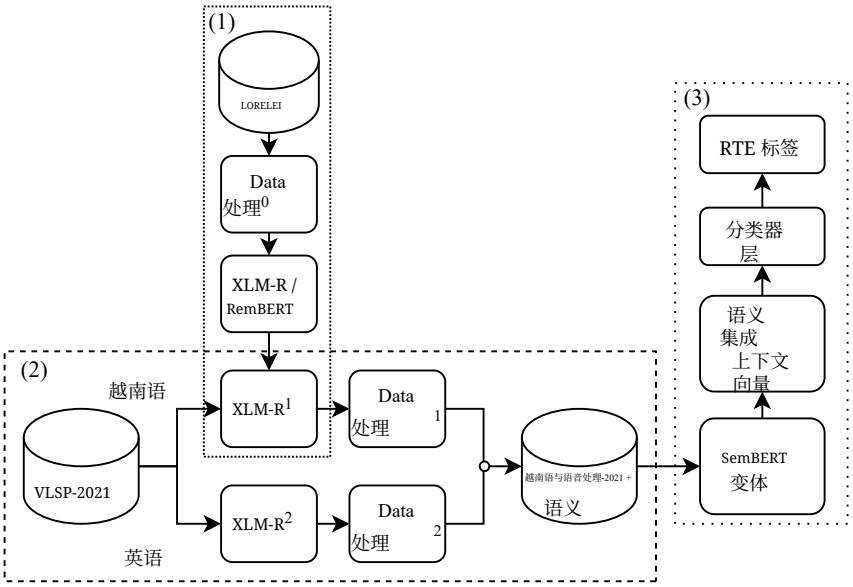

图 2. 实现流程示意图。(1) 针对 SRL任务对 XLM-R 模型进行微调的流程。(2) 对 VLSP-2021语料库进行语义信息抽取的流程。(3) 针对 RTE问题对 SemBERT变体模型进行微调的流程。

我们的实验过程分为三个阶段（图2）：1）针对SRL问题对模型进行微调；2）使用第一阶段微调后的模型为语料库提取语义标签；3）针对RTE问题，对语义认知上下文表示模型进行微调。本文采用一种面向越南语的新方法来处理RTE问题。此外，实验结果表明，SRL的作用可以为在越南语中发展这一问题奠定基础。

II. 相关工作

文本蕴含识别问题吸引了全球自然语言处理领域研究界的关注，值得一提的是，RTE-1 [1] - RTE-7 [2]系列会议自2005年至2011年连续举办。这一系列会议旨在提出有助于提升结果并为这项任务开发数据的方法。然而，越南语文本蕴含识别问题相对较新，但同样吸引了研究界。2021年，越南语与语音处理会议 [3]推出了首个文本蕴含识别数据集。该数据集面临的挑战是同时涵盖越南语和英语。然而，VLSP-2021会议上针对文本蕴含识别问题的方法仍存在局限。

自 Vaswani等人提出 Transformer 架构以来，2017 [4], 一系列模型利用该架构的编码组件学习上下文词表示（BERT [5], XLM-R [6], 和 XLNet [7]），在文本蕴含识别问题上的表现优于传统深度学习方法。同样，PhoBERT [8] 和 BERTweet [9] 等优秀模型也为越南语领域带来了良好的效果。尽管这些模型表现良好，但它们并未考虑输入文本的语义层面。

另一方面，几乎没有研究通过 SRL问题利用语义表示来解决自然语言理解问题，尤其是越南语文本蕴含识别。相比之下，在国外的研究中，Zhang等人提出了 SemBERT 模型[10], ，该模型将 BERT 模型的上下文表示与语义表示相结合，从而在英语和中文自然语言理解任务上的表现比 BERT 模型高约 1%。此外，Shi等人 [11], 和 Yih 等人 [12] 的多项研究也应用了 SRL 问题，并取得了颇有前景的成果。

III. 方法论

A. 语义感知*BERT*

语义感知BERT（SemBERT）模型 [10]由两个组件构成：编码和集成，其中编码组件包含两个模块。第一个模块由一个BERT模型构成，负责学习输入文本的上下文表示。由于BERT模型采用WordPiece词语切分方法，输入文本会被切分为词甚至子词。因此，BERT模型的输出将经过卷积神经网络（CNN）层，以对齐至词级别。同时，第二个模块包含双向门控循环单元模型 [13],，该模型将通过与输入文本相对应的语义标签序列学习语义表示。集成组件则负责将表示上下文和语义表示的两个向量整合在一起，形成一个表示语义感知上下文的向量，用于下游任务的微调。

B. *SemBERT* 变体

我们将 SemBERT架构中的 BERT 模型所使用的上下文学习模块修改为 mBERT 和 XLM-

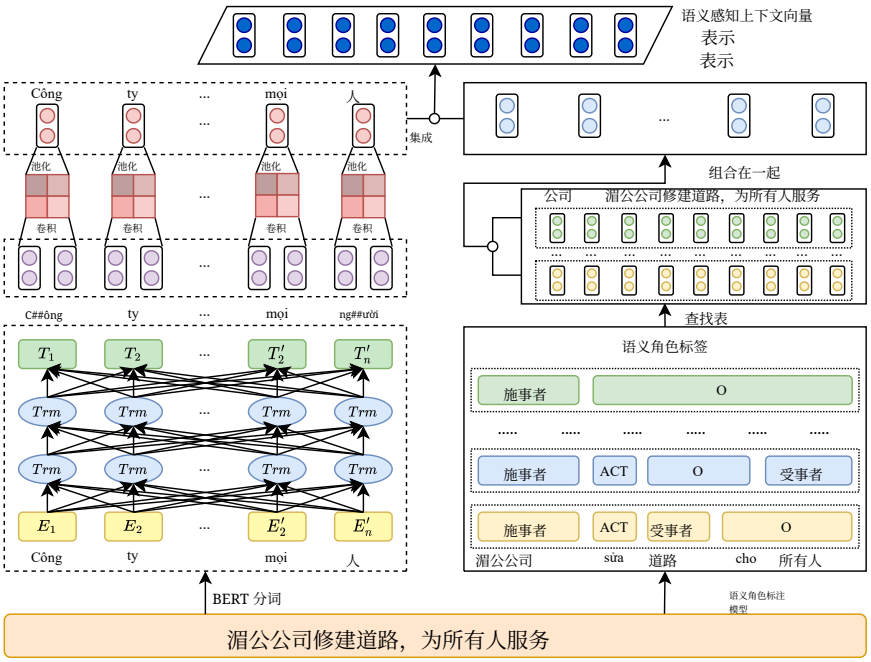

图3. SemBERT模型概览　　该模型，语义角色标注模型专门针对SRL问题进行训练，旨在提取文本输入中的语义标签　　r
。

R 模型。作出这一修改的原因是，实验所选用的语料库包含越南语和英语。另一方面，mBERT 和 XLM-R 模型均已在包含越南语和英语的大型多语言语料库上进行预训练，因此，针对输入语料库学习上下文将比使用 BERT 模型取得更好的结果。

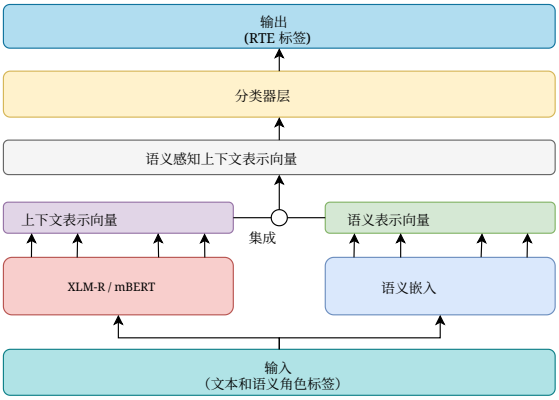

图4. SemBERT变体模型及其微调过程的可视化。

对于英语，已有一些研究提出了用于提取文本语义标签的模型。然而，对于越南语，我们难以获取有关语义角色标注任务的研究成果来执行此任务。因此，为了提取输入文本的语义角色标签，我们针对语义角色标注问题对 XLM-R 和 RemBERT 模型进行了微调。此过程旨在使 SemBERT变体模型学习第 III-A 节所述的语义表示。此任务基于输入的序列标注问题方法解决（He 等人[14],和 Fei 等人 [15]）的 IOB格式。

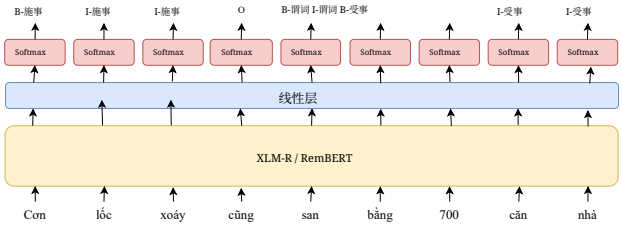

图5. 展示针对SRL任务微调XLM-R/ RemBERT模型的过程。

IV. 实验

训练在 Google Colaboratory 虚拟机上进行。本节将介绍由三项主要任务组成的实现过程：针对 SRL问题对模型进行微调；为输入文本提取语义标签；针对文本蕴含识别问题对 SemBERT 的变体模型进行微调。

A. 数据集

*1) SRL*任务： 我们将 LORELEI 语言包 [16] 中的语料库用于 SRL问题。该问题的数据集包含 1760 个以 XML 文件格式存储的数据点。据作者称，构建该语料库旨在捕捉对句子中发生的事情或情形的基本理解。

2) 文本蕴含识别任务：对于文本蕴含识别问题，我们使用VLSP-2021语料库。该数据集包含 16,185 个训练数据点和 4177 个测试数据点。两个数据集的标签分布均衡，且包含越南语和英语（表 I）。

表 I VLSP-2021 训练集和测试集中越南语与英语数据点的统计信息。

	训练集		测试集	
	文本 *1*_	文本 *2*_	文本 *1*_	文本 *2*_
Vie	8685	16185	2118	4177
Eng	7500	0	2059	0

B. 微调模型以用于语义角色标注

我们针对语义角色标注任务，对 RemBERT 和 XLM-R 模型进行了微调实验。Chung 在 2020 [17], 提出了 RemBERT 模型，这是 XLM-R 模型的改进版本。在这项工作中，Chung 着重研究了改变模型层大小所产生的影响。具体而言，他们发现，增大嵌入层的输出维度能够提升后续任务的微调效果。此外，该模型还在包含多种语言（110 种语言）的大型数据集上进行训练。在微调实体识别任务时，RemBERT 模型取得的结果（F1分数）比 XLM-R 模型高约 5%（此结果直接引用自作者 Chung [17] 的研究）。

在自然语言中，一个句子可以包含多个动词（谓词）。因此，论元对于句子中出现的每个动词都可能承担不同的语义角色。因此，我们采用 k 折交叉验证方法，其中 $k = 10$，训练十个模型，以从句子中提取语义信息的多个方面。所有模型的微调超参数均设置为相同值：学习率 *1e-5*、批量大小 *4*、权重衰减 *0.01*，以及训练轮次 *40*（因为从第 30th 个训练轮次开始，模型性能不再提升）。

表II 显示，在 LORELEI语料库上针对语义角色标注问题进行微调时，XLM-R 模型的性能优于 RemBERT 模型。因此，我们选择微调后的 XLM-R 模型对 VLSP-2021语料库进行语义抽取。

C. 语义标签抽取

VLSP-2021语料库同时包含英语和越南语。因此，英语和越南语的语义标签抽取与数据预处理分别进行。随后，将二者合并，以获得完整的语义信息语料库。

对于越南语，语义抽取过程更加复杂。我们在 LORELEI语料库上针对语义角色标注任务对这十个 XLM-R[1] 模型进行了微调，以预测越南语领域 VLSP-2021语料库的语义标签。接下来，将合并这十个模型的输出数据。之后，我们进入数据预处理阶段，其中包括以下步骤：

- 删除输出数据中的重复数据点。
- 剔除标签序列中同时包含两个动词标签的数据点。
- 合并属于同一句子的语义标签序列。

对于英语，我们使用在 ConLL12语料库上微调的 XLM-R[2] 模型，该语料库是 Oliveira 的工作 [18] 的一部分。在英语领域对 VLSP-2021语料库预测语义标签后，我们进行数据预处理阶段。数据预处理主要检查并删除模型无法预测的数据点。

D. 用于文本蕴含识别的 *SemBERT* 变体模型微调

表示语义表示和上下文表示的两个向量将组合起来，形成一个表示语义认知上下文的向量，h。向量 h将直接输入分类层，以生成对应于同意、不同意和中立标签的预测概率。VLSP-2021 训练语料库将按 8/2 的比例划分为训练集和验证集，并确保两个集合中的标签数量和语言分布均衡。此微调过程的损失函数为交叉熵；AdamW 优化算法及超参数的设置如表III所示。

V. 结果

表 IV 展示了各模型在 VLSP-2021 测试语料库上的性能评估结果。结果表明，通过 SRL 问题整合语义表示与预训练语言模型的上下文表示，能够提升结果。具体而言，在微调 mBERT、XLM-R_{base} 和 XLM-R^{large} 模型时，若结合用于文本蕴含识别问题的语义表示，结果分别为 76.12 %、81.25% 和 88.89%（F_1-分数）。与未整合语义表示的 mBERT、XLM-R_{base} 和 XLM-R^{large}模型相比，这一结果分别提高了 1.42%、0.09% 和 0.98%。

表 V 展示了 VLSP-2021 测试语料库中各语言文本蕴含识别问题的结果。从表中可以看出，mBERT、XLM-R_{base}和 XLM-R^{large} 模型结合语义后，越南语的预测结果优于英语领域，分别达到 76.22%、82.66% 和 89.98%（F_1分数）。具体而言，各模型的结果分别提升了 0.23% 、 2.99% 和 2.26%。另一方面，若考察未融入语义信息的 XLM-R^{large}模型，其英语领域的结果较高，为 87.97%。然而，在结合语义信息后，该模型的表现有所提升，优于原始 XLM-R^{large}模型。这一结果证明，越南语是一种语义丰富的语言，因此丰富越南语文本的上下文有助于提升模型理解自然语言的能力。

A. 错误分析

图6展示了 mBERT、XLM-R_{base}和 XLM-R^{large} 模型在未采用语义匹配时的错误预测趋势

表II 使用K折交叉评估法在LORELEI数据集上对XLM-R和RemBERT两个模型进行微调后的语义角色标注（SRL）任务结果（$k = 10$）（%）。

	XLM-R			RemBERT		
	精确率	召回率	F_1-分数	精确率	召回率	F_1-分数
模型 1	36.72	33.62	35.10	38.44	32.16	35.02
模型 2	44.83	30.37	36.21	37.53	30.19	33.46
模型 3	39.62	31.92	35.36	46.00	28.94	35.53
模型 4	41.34	30.21	34.91	47.67	29.49	36.43
模型 5	38.82	31.72	34.92	35.74	27.62	31.16
模型 6	38.20	32.34	35.03	37.56	28.87	32.64
模型 7	36.72	32.20	34.30	38.58	30.01	33.78
模型 8	41.28	31.42	35.69	37.06	31.15	34.01
模型 9	42.32	33.21	37.22	40.47	30.79	34.97
模型 10	40.91	33.67	36.94	38.26	33.04	35.46
平均值	**40.08**	**32.07**	**35.57**	**39.73**	**30.23**	**34.25**

表 III 用于微调 RTE问题的超参数。

	mBERT	$XLM\text{-}R_{base}$	$XLM\text{-}R^{large}$
学习率	$2e-5$	$2e-5$	$1e-5$
权重衰减	0.01	0.01	0.01
批量大小	12	12	12
最大长度	256	256	256
训练轮次	5	5	5

以及在 VLSP- 2021年各标签上采用语义匹配的模型。可以观察到，三个模型（包括融入语义的 mBERT、$XLM\text{-}R_{base}$ 和 $XLM\text{-}R^{large}$）在中立 标签上的预测效果均较好。具体而言，与未融入语义的模型相比，它们的错误预测率分别降低了 3.95%、4.01% 和 1.79%。然而，融入语义的 mBERT 和 $XLM\text{-}R_{base}$ 模型在不同意 标签上的错误率分别上升了 0.66% 和 2.95%。对于同意标签，融入语义的 mBERT 和 $XLM\text{-}R^{large}$ 模型的错误率分别降低了 1.26% 和 0.76%。在这三个语义关联模型中，$XLM\text{-}R^{large}$ 模型改善了所有三个标签的错误率。

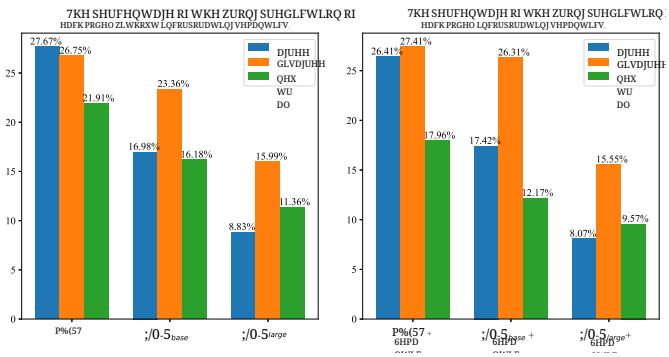

图 6。各模型在每个标签上的错误预测百分比。

B. 谓词数量的影响

谓词或动词是语义角色标注问题中的重要组成部分；谓词可以决定相关成分的语义角色。因此，输入文本中谓词的数量会影响模型性能。基于这一实验策略，我们也开展了实验以得出结论。表VI显示，当谓词数量为2时，结合语义信息的 mBERT 和 XLM-R-大型版模型取得了最佳结果。同时，XLM-R-基础版模型在谓词数量为3时表现最佳。这一结果可以说明，在自然语言中，文本同时包含多个动词的情况较为少见。因此，大量谓词可能会导致语义角色标注模型在语义提取过程中产生混淆。

VI. 结论

我们介绍了文本蕴含识别问题的实现过程和实验结果。对比结果表明，与基础上下文表示模型相比，结合语义表示的 mBERT、$XLM\text{-}R_{base}$ 和 $XLM\text{-}R^{large}$ 模型具有更好的性能。具体而言，具有上下文和语义表示的预训练模型，其性能提升了约 1%。然而，在进行实验时，我们在获取语料库方面遇到了困难，并且语义角色标注问题所使用的语料库规模有限，尤其是 LORELEI语料库。因此，针对语义角色标注问题对 XLM-R 和 RemBERT 模型进行微调的效果并不理想。这项工作影响了越南语 VLSP-2021语料库上的语义标签抽取。因此，针对文本蕴含识别问题对 SemBERT变体模型进行微调也可能受到影响。

从实验结果可以看出，浅层文本表示通过 SRL 任务在自然语言理解中发挥着关键作用。尤其是在越南语领域，目前针对这一任务的语料库研究还不多。基于文本蕴含识别中的 SRL 问题取得的积极结果，为在越南语领域开展这项工作奠定了基础。

表IV VLSP-2021文本蕴含识别问题测试语料库上的预测结果（%）。

模型	准确率	精确率	召回率	F_1-分数
mBERT	74.56	75.02	74.56	74.70
XLM-R_{base}	81.14	81.17	81.16	81.16
XLM-R^{large}	87.92	87.94	87.94	87.91
mBERT + 语义	76.07	76.22	76.07	**76.12**
XLM-R_{base} +语义	81.34	81.38	81.36	**81.25**
XLM-R_{large} +语义	88.91	88.91	88.93	**88.89**

表V VLSP-2021测试语料库中各语言文本蕴含识别问题的预测结果（%）。

	越南语			英语		
	精确率	召回率	F_1-分数	精确率	召回率	F_1-分数
mBERT	75.66	75.33	75.35	74.47	73.78	73.97
XLM-R_{base}	81.42	81.43	81.43	80.87	80.87	80.83
XLM-R^{large}	87.89	87.85	87.86	87.98	88.09	87.97
mBERT + 语义	76.33	76.27	76.22	76.15	75.90	75.99
XLM-R_{base} +语义	82.81	82.72	82.66	79.80	79.90	79.67
XLM-R_{large} +语义	90.04	89.97	89.98	87.68	87.80	87.72

表 VI 纳入模型时，标签序列中谓词数量对实验结果的影响（%）。

模型 谓词数量	1	2	3	4	5
mBERT + 语义	74.83	**76.12**	75.10	75.90	75.77
XLM-R_{base} +语义	80.49	81.01	**81.25**	80.12	80.51
XLM-R_{large} +语义	87.83	**88.89**	88.27	88.21	87.45

致谢

本研究由越南国家大学胡志明市校区（VNU-HCM）资助，资助编号为 DS2022- 26-01。

参考文献

[1] I. Dagan、O. Glickman 和 B. Magnini，“The pascal recognising textual entailment challenge”，载于机器学习挑战赛研讨会。Springer，2005年，第 177–190 页。[2] L. Bentivogli、P. Clark、I. Dagan 和 D. Giampiccolo，“The seventh pascal recognizing textual entailment challenge.” 载于*TAC*。Citeseer，2011年。[3] N. T. Quyen、H. T. Anh、N. T. M. Huyen 和 N. Lien，“VLSP 2021 - vnNLI 挑战赛：越南语及英越文本蕴含”，*VNU* 科学学报：计算机科学与通信工程，第 38 卷，第 1 期，2022年。[4] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力就是你所需要的一切”，神经信息处理系统进展，第 30 卷，2017年。[5] J. Devlin、M.-W. Chang、K. Lee 和 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*，2018年。[6] A. Conneau、K. Khandelwal、N. Goyal、V. Chaudhary、G. Wenzek、F. Guzmán、E. Grave、M. Ott、L. Zettlemoyer 和 V. Stoyanov，“大规模无监督跨语言表示学习”，*arXiv* 预印本 *arXiv:1911.02116*，2019年。[7] Z. Yang、Z. Dai、Y. Yang、J. Carbonell、R. R. Salakhutdinov 和 Q. V. Le，“Xlnet：用于语言理解的广义自回归预训练”，神经信息处理系统进展，第 32 卷，2019年。

[8] D. Q. Nguyen 和 A. T. Nguyen，“PhoBERT：越南语预训练语言模型”，*arXiv* 预印本 *arXiv:2003.00744*，2020年。[9] D. Q. Nguyen、T. Vu 和 A. T. Nguyen，“BERTweet：用于英文推文的预训练语言模型”，*arXiv* 预印本 *arXiv:2005 10200*，2020年。.[10] Z. Zhang、Y. Wu、H. Zhao、Z. Li、S. Zhang、X. Zhou 和 X. Zhou，“用于语言理解的语义感知BERT”，载于人工智能促进协会人工智能会议论文集，第 34 卷，第 05 期，2020年，第 9628–9635 页。[11] C. Shi、S. Liu、S. Ren、S. Feng、M. Li、M. Zhou、X. Sun 和 H. Wang，“用于机器翻译的基于知识的语义嵌入”，载于计算语言学协会第 *54* 届年会论文集（第 *1* 卷：长篇论文），2016年，第 2245–2254 页。[12] W.-t. Yih、M. Richardson、C. Meek、M.-W. Chang 和 J. Suh，“语义解析标注对知识库问答的价值”，载于计算语言学协会第 *54* 届年会论文集（第 *2* 卷：短篇论文），2016年，第 201–206 页。[13] K. Cho、B. Van Merriënboer、D. Bahdanau 和 Y. Bengio，“关于神经机器翻译的特性：编码器-解码器方法”，*arXiv* 预印本 *arXiv:1409.1259*，2014年。[14] L. He、K. Lee、M. Lewis 和 L. Zettlemoyer，“深度语义角色标注：哪些方法有效，下一步是什么”，载于计算语言学协会第 *55* 届年会论文集（第 *1* 卷：长篇论文），2017年，第 473–483 页。[15] H. Fei、M. Zhang、F. Li 和 D. Ji，“通过模型迁移实现跨语言语义角色标注”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，第 2427–2437 页，2020年。[16] S. Strassel 和 J. Tracey，“LORELEI 语言包：低资源语言技术开发所需的数据、工具和资源”，载于第十届国际语言资源与评测会议（*LREC'16*）论文集，2016年，第 3273–3280 页。[17] H. W. Chung、T. Fevry、H. Tsai、M. Johnson 和 S. Ruder，“重新思考预训练语言模型中的嵌入耦合”，*arXiv* 预印本 *arXiv:2010.12821*，2020年。[18] S. Oliveira、D. Loureiro 和 A. Jorge，“利用 Transformer 和迁移学习改进葡萄牙语语义角色标注”，载于*2021 IEEE* 第八届数据科学与高级分析国际会议（*DSAA*）。IEEE，2021年，第 1–9 页。