相似音节增强的越汉跨语言实体消歧方法

李裕娟[1,2]，宋燃[1,2]，毛存礼[*1,2]，黄于欣[1,2]，高盛祥[1,2]，陆杉[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1064406374@qq.com, song_ran@163.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, lushan88d@163.com

摘要

跨语言实体消歧是在源语言句子中找到目标语言相对应的实体，对跨语言自然语言处理任务有重要支撑。现有跨语言实体消歧方法在资源丰富的语言上能得到较好的效果，但在资源稀缺的语种上效果不佳，其中越南语-汉语就是一对典型的低资源语言；另一方面，汉语和越南语是非同源语言存在较大差异，跨语言表征困难；因此现有的方法很难适用于越南语-汉语的实体消歧。事实上，汉语和越南语具有相似的音节特点，能够增强越-汉跨语言的实体表示。为更好的融合音节特征，我们提出相似音节增强的越汉跨语言实体消歧方法，缓解了越南语-汉语数据稀缺和语言差异导致性能不佳。实验表明，所提出方法优于现有的实体消歧方法，在R@1指标下提升了5.63%。

关键词： 实体消歧；音节相似性；越汉跨语言

Similar syllable enhanced cross-lingual entity disambiguation for Vietnamese-Chinese

Yujuan Li[1,2], Ran Song[1,2], Cunli Mao[*1,2], Yuxin Huang[1,2], Shengxiang Gao[1,2], Shan Lu[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
1064406374@qq.com, song_ran@163.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, lushan88d@163.com

Abstract

Cross-lingual entity disambiguation is to find the entity corresponding to the target language in the source language sentence, which is an important support for cross-language natural language processing tasks. Existing cross-lingual entity disambiguation methods can achieve good results in languages with abundant resources, but poor results in languages with scarce resources. Among them, Vietnamese-Chinese is a typical pair of low-resource languages; on the other hand, Chinese and Vietnamese are non-cognate languages with large differences, and cross-lingual representation is difficult; therefore, existing methods are difficult to apply to Vietnamese-Chinese entity disambiguation. In fact, Chinese and Vietnamese share similar syllable features that can enhance entity representation across Vietnamese-Chinese cross-languages. In order to better integrate syllable features, we propose a similar syllable-enhanced Vietnamese-Chinese cross-language entity disambiguation method, which alleviates the poor performance caused by the scarcity of Vietnamese-Chinese data and language differences. Experiments show that the proposed method is superior to existing entity disambiguation methods, and improves by 5.63% under the R@1 index.

*毛存礼(通讯作者):maocunli@163.com

第二十二届中国计算语言学大会论文集，第208页-第219页，哈尔滨，中国，2023年8月3日至5日。

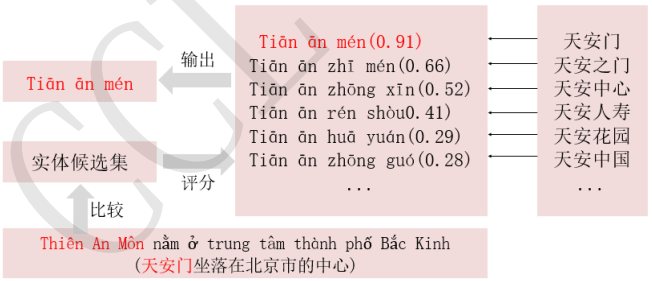
Keywords: Entity Disambiguation , Syllable similarity , Cross-lingual for vietnamese-chinese

1 引言

实体消歧（Entity Disambiguation，ED）的目的是将非结构化文本中提到的实体链接到相应的知识库实体。ED的挑战在于待处理实体的歧义性,例如在文本中提到的“世界杯”和知识库中的实体(如“FIFA世界杯”和“橄榄球世界杯”)存在歧义。ED模型通过建模提及实体的局部信息和文本的全局语义信息，以确定对应的目标实体。提高ED效果的关键在于有效地结合提及信息和上下文信息(Ganea and Hofmann, 2017; Guo and Barbosa, 2018)。

许多跨语言任务依靠多语言知识库提升其性能，如问答(Yang et al., 2017; Yang et al., 2018)、推荐(Cao et al., 2019)和信息抽取(Kumar, 2017)。跨语言实体消歧是够将原文本中提及的实体，在另一语言的知识库中进行匹配，能够为跨语言任务提供支持。目前，跨语言实体消歧主要依赖于多语言预训练模型，多语言预训练模型能够把源语言和目标语言的表示映射到同一语义空间下，以解决跨语言表示的问题并提升跨语言实体消歧效果。得益于预训练语言模型强大的表示能力，富资源语言上跨语言实体消歧已经达到很好的效果，如图1(a)所示，直接使用未经过预训练的跨语言模型计算提及实体和备选实体的相似度就能得到很好的效果。例如，对于提及实体“Washington”，可以在汉语中找到正确的对应实体“乔治·华盛顿”，并且具有高置信度，即目标实体与其他实体的得分差异度较大。

(a) 英中实体消歧　　(b) 越汉实体消歧



(c) 越汉实体消歧

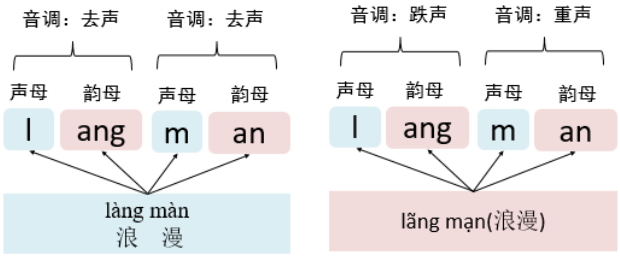
Figure 1: 实体消歧效果对比示意图

然而，大多数多语言预训练模型在越南语上表现不佳。这是因为在M-BERT(Devlin et al., 2018)的训练数据中，英语的数据量是越南语的116倍，导致越南语的表示效果远不如富资源语言，如英语、等。如图1(b)所示，在当利用模型计算越南语实体(Thiên An Môn)和汉语备选实体之间的相似度时，目标实体未被正确匹配，并且不同实体之间的相似度差距较小，置信度较低。因此，依赖于多语言预训练模型进行越汉跨语言实体消歧无法取得良好的效果。分析越南语和汉语之间的关系，并找到越南语-汉语之间的有效特征，能够改善越汉实体消歧的效果。

国家自然科学基金(U21B2027,61972186,62166023,62266028);云南省科技重大专项(202103AA080015,202202AD080003,202202AD080004);云南省基础研究计划项目(202301AS070047,202301AT070471)

事实上，越南语自古受到汉字文化的深远影响(李靖, 2008)，所以越南语的书写、句法以及读音等很多方面仍然保留着许多汉语特色。越南语和汉语类似，构词绝大多数以单音节(或称字)为单位。和多数汉语言以及壮泰语言一样，越南语的音节可由声母、韵母、声调三部分构成(陈雪, 2011)。汉语和越南语的声母中都有爆破音(b、p、t、g、k)和摩擦音(f、s、h)，在韵母中越南语和汉语都有单元音(a、e、i、o、u)和复元音(a、iê、i等)。

如图2，比如“浪漫”这个词，在汉语中拼音为“làng màn”，其声母为“l”(清辅音)，韵母为“ang”(开合中元音+鼻音)，声调为第四声。在越南语中为“lãng mạn”，其声母为“l”(清辅音)，韵母为“ang”(中元音+鼻音)，声调为重声。



(a) 汉语　　(b) 越南语

Figure 2: 汉语拼音和越南语音节分析图

基于以上观察，可以发现越南语和汉语拼音具有相似的音节，但并非所有单词都具有完全相似的结构。例如，“广告”一词在汉语中的拼音为“guǎng gào”，在越南语中的拼音为“quảng cáo”，两者的字符串存在差异。因此，需要将字符串分割成更细粒度字符片段进行比较，以便逐一比较各个字符片段。通过更细粒度的分割，汉语中的“uǎng”字符串和越南语中的“uảng”字符串将显示出更高的相似度，从而提高越南语和汉语拼音之间的整体相似度。因此，将跨语言实体消歧中的汉语和越南语提及的实体的音节结构纳入考虑，可以提高实体消歧的效果。

为了更好融合音节特征并提升汉语实体消歧的性能，本文提出一种相似音节增强的越汉跨语言实体消歧方法(Similar syllable enhanced cross-lingual entity disambiguation for Vietnamese-Chinese，VCED)。如图1(c)所示，首先利用越南语提及实体与汉语实体拼音之间的结构相似性增强实体表示。由于越南语与拼音并不是严格对齐的，因此我们利用N-gram编码对越南语和拼音进行不同字符粒度的切分，并计算他们之间的相似度。然后，使用M-BERT预训练语言模型计算越南语文本与汉语实体描述的语义相似度。最后，综合字符和语义的相似度得到目标实体。实验结果表明，本文提出的方法在越南语-汉语的跨语言实体消歧任务上准确率显著提高。

简而言之，本文工作贡献如下：

(1) 本文分析了越南语与中文之间的语言特点，并找到了音节之间的相似性，并构建了越南语-汉语的跨语言实体消歧数据集，以支持后续实体消歧的研究。

(2) 提出了面向越南语-汉语的音节字符及文本语义表示方法，增强了实体的表示，从而提升越南语-汉语跨语言实体消歧模型的性能。

(3) 我们进行了广泛的实验，结果表明所提出方法可以提升了越南语-汉语实体消歧的性能，效果优于现有的实体消歧模型。

2 相关工作

实体消歧方法分为两类。一种是基于实体特征的实体消歧方法，根据实体和关系的语义特征、实体和关系的上下文特征以及实体出现的频率特征来消除实体歧义。另一种是基于神经网络的实体消歧方法，该方法是利用知识图中的图结构特征，使用神经网络模型进行端到端的实体消歧。

基于实体特征的实体消歧方法提出了一种实体相似度模型来度量模糊实体之间的差异。命名实体消歧系统(Daiber et al., 2013)主要依靠实体上下文相似性度量来消歧。Adjali et al. (2020)使用实体语义相似性、上下文相似性和提及概率来消除实体歧义。MCKR(Hu et al.,

2021)采用多层感知器提取缺失数据与观测数据的交互特征。

Barrena et al. (2015)通过融合来自不同数据源的实体、名称、文本和维基百科信息的概率模型，发现这些特征在实体消歧中具有明显的互补作用。Tsai and Roth (2016)通过用相应的实体标记替换每个实体提及来联合训练单词和实体的单语嵌入，再基于维基百科的语言间链接，使它们学习从多种语言到英语词嵌入空间的投影函数，最后，把源语言实体文本的上下文嵌入同样也投影到英语空间，并与英语实体嵌入进行相似度计算完成实体消歧。Zwicklbauer et al. (2016a)为实体消歧提供了一种实体语义嵌入表示模型，他们使用Word2Vec(Mikolov et al., 2013)方法嵌入实体，并在RDF图上使用随机游走方法构建实体序列。Bouarroudj et al. (2022)提出了一种短文本的命名实体识别，他们使用WordNet进行实体上下文扩展，再基于语义和句法度量对候选词进行排名从而进行实体消歧。Tam et al. (2022)提出了使用知识图谱正则化的条件掩码实体模型（CMEM-KG），其中上下文中的多个提及可以在一次前向传递中消歧。Liu et al. (2022)提出使用KG嵌入进行语义表解释和实体消歧，旨在在先前识别的实体注释之后添加语义消歧步骤，考虑整个列作为上下文，并使用图嵌入来捕获实体之间的潜在关系，以提高它们的消歧性。无论是基于聚类还是基于实体链接，实体与实体、实体与文本、文本与文本之间的相似度计算是实体消歧的核心问题。这些计算方法主要利用自然语言处理技术来提取实体的特征。这些方法虽然取得了较好的性能，但越南语特征可扩展性差，表示能力不足，在实体消歧中容易造成错误传播。

基于神经网络的实体消歧方法采用端到端的机制来提高实体消歧的准确性。除了实体和关系的特征之外，研究人员还利用知识图的图结构特征进一步提高实体消歧的效果。RSJoint(Geng et al., 2021)集成了卷积神经网络和递归神经网络来消除实体歧义和提取关系。Guo and Barbosa (2014)通过估计候选实体的Topic-sensitive PageRank值(Haveliwala, 2002)，结合知识图上的随机游走方法进行实体消歧。Alhelbawy and Gaizauskas (2014)使用基于图的方法进行联合实体消歧，该方法将文本中的所有实体表示为图中的节点，然后根据节点的PageRank值对其进行排序，并根据值的大小进行实体消歧。Singh et al. (2011)利用图形模型来消除文档之间的实体歧义。Zwicklbauer et al. (2016b)设计了一种利用上述实体知识图上的个性化PageRank值的集体消歧方法，该方法依赖于集体链接算法进行实体消歧。

研究人员还尝试使用深度学习方法来消除歧义，并取得了良好的效果。Ganea and Hofmann (2017)采用知识图中的实体嵌入，采用基于注意的方法获得嵌入向量，并考虑实体之间的相干性进行联合消歧。与依赖监督或启发式方法预测实体关系不同，Le and Titov (2018)将实体关系作为神经实体链接模型中的隐变量，以端到端的机制实现实体消歧。DeepType(Raiman and Raiman, 2018)通过将一个符号特征和一个典型特征结合到神经网络推理中，解决了实体消歧问题。研究者设想了一种类型模型，并利用它来限制网络的输出以适应结构特征。他们提出了一种两阶段的实体消歧算法，首先创建一个类型系统，然后用它来训练神经网络。Hu et al. (2020)和Grover and Leskovec (2016)都是使用图神经网络模型解决实体消歧的问题。Phan et al. (2017)提出了一种深度神经网络方法NeuPL来计算实体之间的语义相似度。NeuPL是第一个使用长短期记忆网络消除实体歧义的模型。但基于神经网络的方法的局限性在于缺乏很好的解释性，且越南语的知识图较为稀疏，使用此方法并不能是越汉实体消歧达到很好的效果。

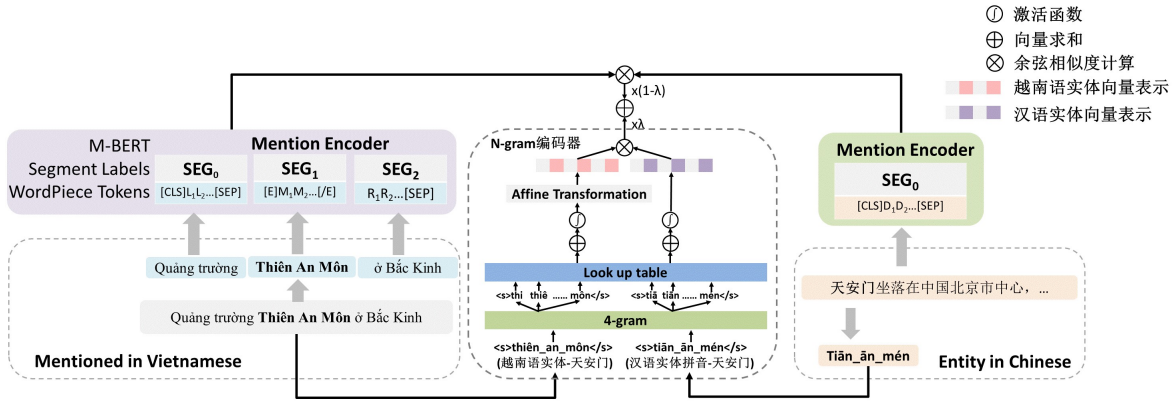


Figure 3: 相似音节增强的越汉跨语言实体消歧模型图

3 方法

我们方法的关键思想是融合越南语提及实体与汉语拼音之间的相似性和文本的相似性来计算实体自身之间的相似度。具体的**模型框架如图3所示**，主要由2个模块组成:(1)文本语义编码，主要是计算越南语实体提及文本和汉语实体描述的相似度；(2)音节字符编码，主要是用于计算越南语实体本身以及汉语实体本身的相似度。

3.1 文本语义编码

3.1.1 越南语文本编码

越南语文本编码是将越南语一段文本中字词序列和分句标签序列融合编码。给定模型输入的越南语文本表示T_{vi}，如图3中左下越南语"Quảng trường Thiên An Môn ở Bắc Kinh."和其中的越南语实体提及表示M_{vi}，如图3中"thiên_an_môn"。模型首先使用WordPiece分别切分越南语实体提及M_{vi}本身以及其所在文本T_{vi}的上下文文本得到对应的子词序列。再使用特殊的分隔符"[E]"和"[\E]"对实体提及和上下文文本进行标记，得到最终的子词序列，具体如图3所示。在送入多语言预训练模型M-BERT前，使用分句标签来对M_{vi}以及上下文文本进行标记，其中上文子词序列对应标签为"0"，实体提及子词序列对应标签为"1"，下文子词序列对应标签为"2"，得到最终的分句标签序列。最后把所得子词序列以及分句标签送入M-BERT模型得到最终的向量表示为公式1:

$$M_{vi}^{c} = M\text{-}BERT(subs_{vi} + seg_{vi}) \tag{1}$$

其中$subs_{vi}$为最终的子词序列，seg_{vi}为最终的分句标签序列。M_{vi}^{c}是M_{vi}的上下文向量表示。

3.1.2 汉语文本编码

汉语文本编码是将一段汉语文本中字词序列和分句标签序列融合编码。给定模型输入的汉语文本表示T_{zh}，该文本为实体在维基百科页面下介绍的第一句话，如图3中右下汉语"天安门坐落在北中国北京市中心..."和其中的汉语实体拼音字符表示E_{zh}。首先使用WordPiece切分该文本得到子词序列$subs_{zh}$。再使用分句标签"0"标记该子词序列得到分句标签序列seg_{zh}后，通过M-BERT得到汉语实体E_{zh} 的上下文向量表示E_{vi}^{c}，具体公式如2所示:

$$E_{zh}^{c} = M\text{-}BERT(subs_{zh} + seg_{zh}) \tag{2}$$

3.2 音节字符编码

N-gram能有效获取上下文的信息，因此本文通过使用N-gram模型丰富汉语以及越南语实体的编码向量表示。首先，将字符长度为m的字符串表示为字符序列$X = [x_1, x_2, ..., x_m]$,其中字符包括空格字符（空格以"_"代替）以及特殊的起始符和结束符，例如图3中越南语实体"thiên_an_môn"表示为字符序列["<s>"，"t"，"h",...,"n"，"</s>"]。为了获取字符串的多粒度表示，首先对越南语实体提及和汉语实体拼音表示进行多粒度4元切分，例如越南语实体"thiên_an_môn"，切分为"<s>thi"、"thiê"、"hiên"、"iên_"、"ên_a"、"n_an"、"_an_"、"an_m"、"n_mô"、"_môn"以及"môn</s>"共计11个4元组。

给定x_i^j来表示从位置i到位置j的子序列，例如$x_i^j = [x_i, x_{i+1}, ..., x_j]$。字符串$X$的4元嵌入表示$V_n$具体公示如3所示:

$$V_n = \left(\sum_{i=1}^{m} \mathrm{II}(x_i^j \in V) W_{x_i^j} \right) (j = i + n - 1, j \leq m + 2) \tag{3}$$

其中n是预定义的n-gram窗口大小，这里n=4。$m+2$表示字符串长度加上起始符和结束符的长度。$W \in R^{|V| \times d}$是嵌入矩阵，$W_{x_i^j} \in R^d$是x_i^j的向量嵌入表示。V是训练数据集中字符的所有4元组合，II()是指示函数，就是如果一个4元字符组合不在V中则丢弃。

给定汉语实体拼音表示E_{zh}^{py}以及越南语实体提及表示M_{vi}^{n}，通过上述过程得到最终的汉语实体以及越南语实体提及向量表示E_{zh}^{ngram}和E_{vi}^{ngram}，具体公式如4和5所示:

$$E_{zh}^{ngram} = N\text{-}gram(E_{zh}^{py}) \tag{4}$$

$$E_{vi}^{ngram} = N\text{-}gram(M_{vi}^{n}) \tag{5}$$

3.3 相似度计算

最后，将越南语文本提及上下文和汉语实体描述进行相似度计算以及越南语提及实体和汉语实体拼音进行相似度计算。通过以上过程获取得到越南语实体提及的上下文向量表示M_{vi}和自身向量表示E_{vi}^{ngram}，汉语实体描述文本向量表示E_{zh}^{c}和自身向量表示E_{zh}^{ngram}。为了综合考虑上下文全局信息以及自身局部信息，分别对全局信息向量M_{vi}^{c}和E_{zh}^{c}以及局部信息向量E_{vi}^{ngram}和E_{zh}^{ngram}进行相似度计算，具体公式如6和7所示：

$$sim_c = \frac{(M_{vi}^{c} E_{zh}^{c})^T}{||M_{vi}^{c}|| \times ||E_{zh}^{c}||} \tag{6}$$

$$sim_n = \frac{(E_{vi}^{ngram} E_{zh}^{ngram})^T}{||E_{vi}^{ngram}|| \times ||E_{zh}^{ngram}||} \tag{7}$$

在获得全局信息相似度得分sim_c以及局部信息相似度得分sim_n后，把两组相似度得分通过超参数λ进行线性组合，得到包含了上下文信息以及实体自身信息的综合相似度sim_{comb}，具体计算公式如8所示：

$$sim_{comb} = \lambda sim_c + (1-\lambda) sim_n \tag{8}$$

3.4 损失函数

本文选用HingeEmbeddingLoss损失来做为训练所用的目标损失函数。模型训练Loss的计算公式如9所示：

$$Loss = \begin{cases} sim_{comb}^{i} & if y_i = 1 \\ max(0, margin - sim_{comb}^{i}) & if y_i = 0 \end{cases} \tag{9}$$

其中，y_i表示第i对数据的真实标签，若越南语实体提及与中文实体是对齐实体则为1，否则为0。sim_{comb}^{i}表示第i对数据之间的距离，$margin$为两元素之间间隔距离允许的边界值，模型中设置为1。

4 实验

为了验证所提出方法的有效性，首先介绍自构数据集(4.1)和实验结果评估方法(4.2)，然后在常规的实验设置(4.3)下与其他实体消歧方法进行比较，还进行了消融实验(4.4)，以评估不同模块对实体消歧效果的影响和不同预训练语言模型对文本语义模块的影响。最后通过样例分析(4.5)进一步说明了本文提出的方法优于其他模型。

4.1 数据集

WikiANN(Pan et al., 2017)[0]是基于维基百科文章的跨语言命名实体识别和实体链接数据集，其中越南语数据共包含110,535条。我们基于WiKiANN以及维基百科[1]中的多语言链接来构建越汉跨语言实体消歧越南语部分的数据集，之后，再通过维基百科获取越南语相对应的汉语实体以及对应的描述。

数据集	总数量	正例数量	越南语实体数量	未在训练集出现过数量
训练集	60,000	10,000	6,303	—
验证集	10,000	2,000	1,515	755
测试集	11,825	2,365	1,737	801

Table 1: 越汉跨语言实体消歧数据集

数据集共计正例14,365条，划分为训练集、验证集和测试集分别为10,000条、2,000条以及2,365条。为了保证模型能学习到更好的知识，防止模型过拟合，增强模型鲁棒性，故采用交

[0]https://elisa-ie.github.io/wikiann/
[1]https://zh.wikipedia.org/wiki/

错匹配的方式构造出负例数据，其中正负比例为1:4，最终训练集、验证集和测试集的数据数量如表1所示，训练集、验证集和测试集中分别存在6,303、1,515和1,737个不相同的越南语实体提及。而验证集和测试集中分别有755、801个越南语实体提及没有出现在训练集中，分别占比为49.83%和46.11%。

4.2 评价指标

本文的实验选用召回率R@1来对实验结果进行评价，因为对于最终的系统比较，标准的做法是使用top检索实体的准确性(R@1)，R@1的具体计算公式如公式10所示：

$$R@1 = \frac{\sum_{i}^{PT} ||C(M_i^{vi})||}{PT} \tag{10}$$

PT为测试集中正例的数量，$C(M_i^{vi})$为该方法在测试集中越南语实体提及M_i^{vi}的正例和负例的结果，若全部正确则取1，否则取0。

Parameter	Value
BERT embedding size	768
BERT model	M-BERT
n-gram embedding size	100
dropout	0.3
Optimizer	Adam
Learning rate	5e-5
n-gram size	4
layers	9-12

Table 2: 模型主要超参数

4.3 参数设置

本文实验使用Pytorch1.7.1版本进行，选用Adam作为优化器，Batch大小设置为32，Epoch设置为20。学习率调整采用等间隔调整策略，每训练10轮调整学习率为当前学习率的百分之十。为了防止实验过拟合，在部分地方使用Dropout技术.所有实验均在一张RTX3090 Ti上训练，实验的主要超参数指标如表2所示。

4.4 实验结果及分析

4.4.1 对比实验

我们将$VCED$与3个基线模型进行了比较，3个基线都是基于实体特征的实体消歧方法，我们采用召回率R@1作为评价指标，R@1越高表示性能越好。

$WikiME$(Tsai and Roth, 2016)在2016年提出的通过联合训练单词和维基百科标题的多语言嵌入来将非英语文档中的实体提及内容与英语维基百科条目联系起来的模型。

$XELMS_{joint}$(Upadhyay et al., 2018)在2018年提出的一个结合多种语言监督的实体消歧方法。

$ModelF$(Botha et al., 2020)提出训练的双编码器模型，在先前工作的基础上改进了特征表示、负挖掘和辅助实体配对，以提高在低资源条件下实体消歧模型的性能。

通过实验结果可以发现，本文的方法可以有效提升越汉跨语言实体消歧模型的性能，效果优于其他几种对比方法。表3显示出$VCED$和$ModelF$均优于对比方法中的$WikiME$和$XELMS_{joint}$模型，主要是$VCED$和$ModelF$都使用了基于大规模语料训练的预训练多语言模型M-BERT，因此模型初始时就富含汉语和越南语的丰富语言知识，针对

模型	R@1(%)	未见实体R@1(%)
$WikiME$	35.05	16.12
$XELMS_{joint}$	38.69	17.67
$ModelF$	42.22	19.34
VCED	**47.85**	**21.96**

Table 3: 不同模型实验结果

低资源语言的模型训练来说预训练语言模型能提高模型性能。我们的模型在R@1指标下高于$ModelF$模型5.63%，虽然$ModelF$也使用M-BERT预训练语言模型充分地关注上下文语义信息，得到了更好的上下文特征表示向量，但在实体消歧任务中，除了上下文信息以外，实体自身的表示也同样重要，本文的方法加入了实体的音节字符信息，将汉语和越南语实体自身各粒度n元组进行编码从而丰富了实体自身的局部信息，提高了模型效果。

我们还在未见实体上做了模型性能测试，实验结果表明，VCED模型在未见实体上的性能也高于其他几个基线模型，这是因为预训练语言模型在没有见过实体的上下文表示比较弱，而相比较于Model F模型VCED模型加入了实体自身的表示，提高了模型的效果。

4.4.2 消融实验

现在预训练语言模型多种多样，我们对比了召回率与不同预训练语言模型之间的关系，在M-BERT，LaBSE(Feng et al., 2020)，SBERT(Reimers and Gurevych, 2019)预训练语言模型上做了实验。然后为了测试我们方法在两个模块上对实验结果影响，我们将音节字符编码模块和文本语义编码模块分别进行实验。

从表4可以看出，我们的模型使用的M-BERT预训练语言模型效果高于其他预训练语言模型，这是因为M-BERT富含了丰富的语义信息，所以使用含有丰富的语义信息的预训练语言模型在一定程度上可以提升实体消歧模型的性能。从单独用音节字符编码模块和单独用文本语义编码模块的实验结果来看，单独考虑文本语义信息或单独考虑音节字符信息去提升实体消歧模型性能是远远不够的，需要同时考虑文本语义信息和音节字符信息才能有效提升实体消歧的效果。

预训练语言模型	R@1(%)
$VCED_{onlyM-BERT}$	15.35
$VCED_{onlyN-gram}$	46.13
$VCED_{LaBSE}$	46.19
$VCED_{SBERT}$	45.53
VCED	**47.85**

Table 4: 消融实验

4.4.3 不同λ值实验

为了研究我们方法的召回率与超参数λ的大小之间关系，实验在超参数λ设置为0.1，0.3，0.5，0.7以及0.9时的准确率。实验结果如表5所示:

从表5可以看出超参数λ的大小对实验结果有着显著的影响。通过上述实验可以得知，超参数λ从0.1增长到0.5时，实验结果的召回率随着其增长也增长，λ到0.5时实验召回率达到最高的47.85%，但当λ从0.5继续增长到0.9时，实验结果的召回率随着其增长开始逐渐降低，此组实验证明了实体本身的信息和实体上下文信息在实体消歧任务中同样重要。λ从0.1到0.5，召回率提升了13.52%，而λ从0.5到0.9，召回率只下降了5.98%，从此结果可以看出上下文全局信息的

λ	R@1(%)
0.1	34.33
0.3	40.58
0.7	43.46
0.9	41.87
0.5	**47.85**

Table 5: 不同 λ 值实验结果

重要程度要高于实体自身，只有合理考虑上下文信息和实体自身信息才能时模型性能达到最好。

4.4.4 音节相似比例实验

为了验证测试集中音节相似比例对模型效果的影响，我们将测试集中数据的音节相似分为4种比例([0-0.25)，[0.25-0.5)，[0.5-0.75)，[0.75-1])，分别测试音节相似比例对模型准确率的影响，其中，值越小表示两个音节之间字符串的相似度越接近。由图4可以看出，随着音节相似度的降低，模型的准确率在不断的提高，这是因为在训练集中音节相似的数据占比较少，训练时模型对于音节相似的实体没有得到足够的学习所以模型的准确率较低，而音节不相似时模型的准确率较高是因为我们提出了4-gram编码器，对于较长的实体音节有更多组合的可能,丰富了音节的表示，从而提高了模型的准确率。

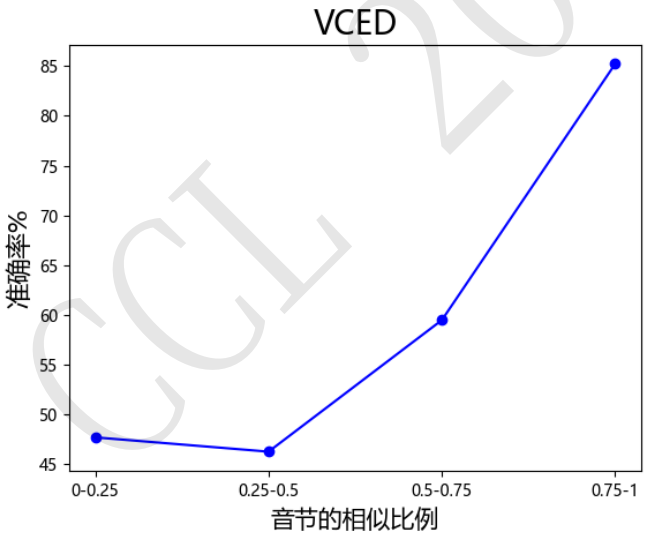


Figure 4: 不同比例的相似音节对模型准确率的影响

4.4.5 数据正负比例实验

为了验证我们方法的有效性并研究在构造数据集时正例与负例比例对实验结果的影响，我们分别在数据集中正负比例分别在1:2，1:3，1:4以及1:5时的结果进行对比，记录每组实验的R@1值，实验结果如表6所示。

从表6中可以看出，数据集中正负比例为1:2时，实验结果的召回率最低只有42.41%。随着负例数量上升到1:4时，实验的召回率随着负例的增加而增加，达到最高的47.85%，而当正负比例到1:5时，实验效果开始下降，可知，在原始数据集只有正例且正例数量固定的情况下，适当增加负例能帮助模型学习到更多的知识，能增加模型的鲁棒性，防止模型过拟合，但当负例达到某个数量时，再靠增加负例来使数据集数量变大提升模型效果是不可行的，因为负利数量过

正负比例	R@1(%)
1:2	42.41
1:3	45.88
1:5	46.13
1:4	**47.85**

Table 6: 不同正负比例实验结果

大会影响模型对正例的敏感度降低模型整体性能，还会无谓的增加训练时间。

4.5 样例分析

我们给出了越南语-汉语实体消歧的实例。基于实体特征的实体消歧模型在越汉实体消歧时，会出现目标实体未被找到的现象。如表7，越南语文本中的实体为“Khonkaen”，由于越南语是低资源语言这一事实，没有较多的文本语义信息，会导致实体消歧模型未找到汉语中相对应的实体“孔敬市”，而利用文本语义信息和音节字符信息结合，有效的提高了实体消歧的准确率。此外，仅仅基于实体的文本语义的方法($ModelF$)无法较好的识别语义信息较弱的越南语实体，而本文提出的方法结合了文本语义信息和音节字符信息，缓解了越南语-汉语数据稀缺导致性能不佳的问题，提高了实体消歧的准确率。第三行是VCED模型top1召回失败样例，原因是越南文本中出现了将军(Tướng)这个词，且提及实体和对应实体“阮知方(ruan zhi fang)”的音节至多只有一个字符是相似的，因此错误的将“阮知方军团”判断成消歧正确的实体。

越南语文本	提及实体	$ModelF$	$VCED$
Sau đó Ratchasuphawadi chuyển tới **Khonkaen**	Khonkaen	孔敬,孔敬大学, 孔敬总领事馆, **孔敬市**,孔敬机场	孔敬市, 孔敬总领事馆,孔敬, 孔敬大学,孔敬机场
Bà chết cùng ngày với **Lê Cung Hoàng**	Lê Cung Hoàng	刘恭煌,黎椿,**黎恭皇**	黎恭皇,黎椿,刘恭煌
Tướng **Nguyễn Tri Phương** bị trọng thương	Nguyễn Tri Phương	**阮知方军团**,阮知方, 阮知方路,阮知方街道, 阮知方军团酒店	**阮知方军团**,阮知方, 阮知方路,阮知方街道, 阮知方军团酒店

Table 7: 越汉跨语言实体消歧样例分析

5 结束语

针对现有实体消歧模型因越汉数据稀缺导致性能不佳的问题，本文提出了一个新的越汉实体消歧模型(VCED)，利用越南语和汉语文本语义信息和音节字符信息增强实体的表示，从而拉近两种语言实体在向量空间上的距离。实验结果表明本文方法的有效性和优越性，在相同的数据集下比其他实体消歧模型提升了5.63%。在之后的工作中，我们考虑融入知识图的图结构特征以提高实体消歧的效果。

参考文献

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *Advances in Information Retrieval: 42nd European Conference on IR*

Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I, pages 463–478. Springer.

Ayman Alhelbawy and Robert Gaizauskas. 2014. Graph ranking for collective named entity disambiguation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 75–80.

Ander Barrena, Aitor Soroa, and Eneko Agirre. 2015. Combining mention context and hyperlinks from wikipedia for named entity disambiguation. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 101–105.

Jan A Botha, Zifei Shan, and Daniel Gillick. 2020. Entity linking in 100 languages. *arXiv preprint arXiv:2011.02690*.

Wissem Bouarroudj, Zizette Boufaida, and Ladjel Bellatreche. 2022. Named entity disambiguation in short texts over knowledge graphs. *Knowledge and Information Systems*, 64(2):325–351.

Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *The world wide web conference*, pages 151–161.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th international conference on semantic systems*, pages 121–124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*.

Zhiqiang Geng, Yanhui Zhang, and Yongming Han. 2021. Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, 429:132–140.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Zhaochen Guo and Denilson Barbosa. 2014. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 499–508.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Taher H Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526.

Linmei Hu, Jiayu Ding, Chuan Shi, Chao Shao, and Shaohua Li. 2020. Graph neural entity disambiguation. *Knowledge-Based Systems*, 195:105620.

Xuan Hu, Yongming Han, and Zhiqiang Geng. 2021. A novel matrix completion model based on the multi-layer perceptron integrating kernel regularization. *IEEE Access*, 9:67042–67050.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645*.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. *arXiv preprint arXiv:1804.10637*.

Jixiong Liu, Viet-Phi Huynh, Yoan Chabot, and Raphael Troncy. 2022. Radar station: Using kg embeddings for semantic table interpretation and entity disambiguation. In *The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings*, pages 498–515. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Minh C Phan, Aixin Sun, Yi Tay, Jialong Han, and Chenliang Li. 2017. Neupl: Attention-based semantic matching and pair-linking for entity disambiguation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1667–1676.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models.

Zhi-Rui Tam, Yi-Lun Wu, and Hong-Han Shuai. 2022. Improving entity disambiguation using knowledge graph regularization. In *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part I*, pages 341–353. Springer.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. *arXiv preprint arXiv:1809.07657*.

Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. 2017. Efficiently answering technical questions—a knowledge graph approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016a. Doser-a knowledge-base-agnostic framework for entity disambiguation using semantic embeddings. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Proceedings 13*, pages 182–198. Springer.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016b. Robust and collective entity disambiguation through semantic embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 425–434.

李靖. 2008. 汉语对越南语的影响. 时代人物, (7):120–121.

陈雪. 2011. 汉语拼音与越南语音节结构之对比研究. 新课程学习*(*下*)*.