DOI:10.20033/j.1003-7241.(2026)04-0129-05

基于语言模型的翻译文本隐形词汇语义自动标注算法研究

刘海琦, 刘淑波

(海军大连舰艇学院 基础部,辽宁 大连 116018)

摘　要:针对当前文本翻译中隐形词汇语义翻译与检测准确率低的问题,旨在提升翻译文本中隐形词汇的翻译准确率,为整体翻译准确率提供依据。研究将 Transformer 语言模型与自动标注算法结合,设计了适用于翻译文本隐形词汇语义的自动标注算法,通过 Transformer 模型对翻译文本建模并构建隐形词汇语义数据库,再利用自动标注算法完成隐形词汇的预测、标注、语义匹配与概率排序,同时简要说明算法在翻译软件中的应用流程及相关计算方法。为验证算法性能,将其与 3 种对比算法开展模拟实验,结果显示该算法标注准确率达 98.5%、误差率 0.2%、Kappa 系数 0.93,计算速度与稳定性均显著优于对比算法;将其应用于翻译软件优化后,软件词汇、语法等翻译指标准确率大幅提升,隐形词汇语义翻译准确率达 98.9%,BLEU 评分提升至 0.92,翻译耗时显著降低。研究证实该算法能有效实现隐形词汇的标注与准确翻译,提升整体翻译效果,但目前仅验证了中英文翻译场景,其对其他语言的适用性仍需进一步研究。

关键词:翻译文本;Transformer;模型;自动标注算法;隐形词汇语义;语言模型

中图分类号: TP391.2　　**文献标志码**: A　　**文章编号**: 1003-7241(2026)04-0129-05

Research on automatic labelling algorithm for invisible lexical semantics of translated texts based on language models

LIU Haiqi, LIU Shubo

(Department of Basics, Dalian Naval Academy, Dalian 116018, Liaoning, China)

Abstract: Aiming at the low accuracy of semantic translation and detection of implicit words in current text translation, this study combines the Transformer language model with the automatic annotation algorithm to design an automatic annotation algorithm suitable for the semantics of implicit words in translated texts. The Transformer model is used to model translated texts and construct an implicit word semantic database, and then the automatic annotation algorithm is utilized to complete the prediction, annotation, semantic matching and probability ranking of implicit words. Meanwhile, the application process and related calculation methods of the algorithm in translation software are briefly explained. Simulation experiments are conducted with three comparison algorithms show that the proposed algorithm achieves an annotation accuracy of 98.5%, an error rate of 0.2% and a Kappa coefficient of 0.93, and its calculation speed and stability are significantly superior to those of the comparison algorithms. After applying it to the optimization of translation software, the accuracy of the software's translation indicators such as vocabulary and grammar is greatly improved, the semantic translation accuracy of implicit words reaches 98.9%, the BLEU score is increased to 0.92, and the translation time is significantly reduced. The study confirms that the algorithm can effectively realize the annotation and accurate translation of implicit words, but it has only been verified in Chinese-English translation scenarios, and its applicability to other languages needs further research.

Keywords: translated text; Transformer; model; automatic labelling algorithm; invisible lexical semantics; language model

随着社会的不断发展,人们生活质量的不断提升,各个国家之间的交流越来越多,但是不同国家之间由于语言不同,交流还存在很多困难[1]。针对交流困难的问题,现在已经有很多有关文本翻译的方法。例如,Abidin 团队为了将楠榜语文本方言 Nyo 翻译成印尼语,提出了直接机器翻译法和统计机器翻译法,对两种方法进行实验,结果显示,直接机器翻译法翻译准确率为 39.32%,统计机器翻译法的翻译准确率为 59.85%[2]。在国内,李舒淇等为了提高译员在执行翻译操作中的翻译效率,利用智能算法平台设计了一种多平台协同翻译系统,对基于智能算法的翻译系统进行实验测试,结果显示,该翻译系统能够提高文本翻译的准确率和效率[3]。此外,针对现有的翻译工具性能不稳定的缺点,刘金硕等提出了一种两级并行翻译方法,对该方法进行实验测试,结果显示,该方法能够提高翻

收稿日期:2025-01-13;**录用日期**:2025-01-29

基金项目:军队院校英语教学联席会第二批教学改革重点研究项目(WY2024017A);军队院校英语教学联席会第二批教学改革一般研究项目(WY2024063B)

作者简介:刘海琦(1981—),女,教授,硕士,研究方向:英语教学、翻译。

引用本文:刘海琦, 刘淑波. 基于语言模型的翻译文本隐形词汇语义自动标注算法研究[J]. 自动化技术与应用, 2026,45(4):129-132, 139. (LIU Haiqi, LIU Shubo. Research on automatic labelling algorithm for invisible lexical semantics of translated texts based on language models0[J]. Techniques of Automation and Applications, 2026,45(4):129-132, 139.)

译速度,降低翻译时间[4]。但是目前的翻译方法对隐形词汇语义的翻译还不够准确,并且对隐形词汇的检测准确率也不高,从而导致文本翻译的准确率低、翻译后的语义出现错误。因此,设计一种能够对翻译文本隐形词汇语义进行标注同时进行翻译的方法显得尤为重要。Transformer 语言模型是一种基础预训练语言模型[5]。而自动标注算法可以对需要的关键词进行自动标注和分类[6]。研究引入该模型通过自注意力机制来表示隐形词汇和多个语义之间的关系,并形成隐形词汇语义数据库。针对翻译文本中的隐形词汇再利用自动标注算法进行自动标注和分类,再对隐形词汇进行翻译。研究旨在提高翻译文本中隐形词汇的翻译准确率,为翻译文本整体翻译的准确率提供依据。研究的创新之处在于,将 Transformer 语言模型和自动标注算法进行结合,并将结合的算法首次用于翻译文本的隐形词汇语义的自动标注和翻译中,以提高翻译效率和准确率。

1 自动标注算法的设计

1.1 自动标注算法

随着社会的不断发展,各个国家之间的交流越加频繁,翻译软件的应用得到了推广[7]。但目前的翻译软件对于文本隐形词汇语义的翻译还存在困难[8-9]。所以设计一个对文本隐形词汇语义进行自动标注的算法以提高翻译软件的翻译准确率是一项亟须解决的问题[10]。语言模型在语音识别、文本翻译、信息检索等方面有广泛的应用[11]。研究选择具有语义完备性、命名识别能力强的 Transformer 语言模型作为研究对象[12]。Transformer 模型包含了编码器和解码器两个部分,自注意力机制是该模型的关键部分[13-14]。该机制的输出表达式如式(1)所示。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

式中,$\boldsymbol{Q}$ 表示查询向量,$\boldsymbol{K}$ 代表键向量矩阵,$\boldsymbol{V}$ 表示数值向量。且 $\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_Q$,$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_K$,$\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_V$,$\boldsymbol{W}$ 表示一个训练矩阵,$\boldsymbol{X}$ 代表输入矩阵,$\boldsymbol{K}^{\mathrm{T}}$ 表示 $\boldsymbol{K}$ 矩阵的转置矩阵,d_k 表示 $\boldsymbol{Q}$,$\boldsymbol{K}$ 矩阵的列数即向量维度。其中编码器是由多头注意力机制,添加和规范,前馈组成,添加和规范层的计算方式如式(2)所示。

$$Y = L(A + M(A))L(A + F(A)) \tag{2}$$

式中,A 表示该层的输入,$M(A)$,$F(A)$ 表示输出。该层的添加是一种残差连接,一般用于解决多层网络训练问题,让网络的注意力全部集中在当前差异的部分。该层的规范是指层规范化,以加快模型收敛速度。前馈层是一个两层的全连接层。由上面描述的三部分组成编码块,多个编码块组成编码器,通过编码器将文本信息转化为编码信息矩阵。Transformer 语言模型的解码器包含了两个多头注意力机制,和一个激活函数层。第 1 个多头注意力机制采用了 Masked 操作,第 2 个多头注意力机制用于计算编码信息矩阵。最后使用 softmax 预测输出。语言模型可以对训练集中的翻译文本和文本隐形词汇进行建模,建立文本隐形词汇和文本语义之间的关系。而基于 Transformer 语言模型的自动批注算法可以根据建立的语言模型对翻译文本的隐形词汇和语义进行预测。基于 Transformer 语言模型的自动标注算法的流程如图 1 所示。

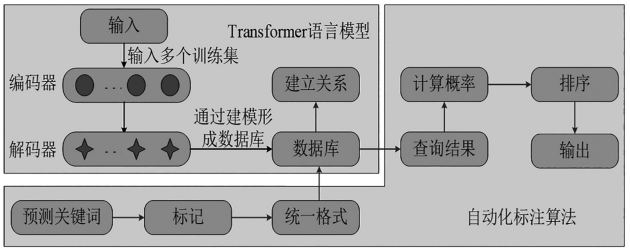


图 1 基于 Transformer 语言模型的自动标注算法的流程

Fig. 1 Flowchart of the automatic annotation algorithm based on the Transformer language model

由图 1 可得,该算法先通过 Transformer 语言模型中的编码器、解码器对翻译文本进行建模,得到一个关于语言翻译的文本数据库,且在语言模型中建立需解决问题之间的关系。之后使用自动标注算法对该问题进行预测。自动标注算法需要引入对解决问题的预测,然后对该预测结果在语言模型建立的模型进行查找,最后输出查找结果[15-16]。自动标注算法的步骤如下所示。

步骤 1 对输入的信息的关键词进行预测并将预测信息进行标记。

步骤 2 将标记的关键词信息格式进行转换。

步骤 3 将关键词信息输入到数据库中对其可能的结果进行查询。

步骤 4 通过查询得到关键词与数据库中不同特征词之间的关系。

步骤 5 对关键词的不同特征词出现的概率进行计算,并按照出现概率的高低进行排序。

步骤 6 将关键词语与特征词之间的可能关系按照降序的方式进行输出。

1.2 自动标注算法应用

将基于 Transformer 语言模型的自动标注算法用于翻译文本隐形词汇语义的检测中,以期提高翻译软件对文本翻译的准确率。基于该算法的翻译软件的工作流程如图 2 所示。

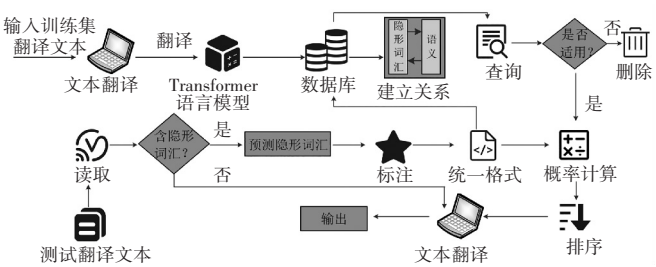


图 2 基于自动标注算法的翻译软件的工作流程

Fig. 2 Workflow of the translation software based on the automatic annotation algorithm

由图 2 可知,翻译软件在工作时,在训练模型中先多批次输入需要翻译的文本,通过算法的 Transformer 语言模型将文本进行翻译,并将文本和翻译结果一起存储在数据库中,并对其进行建模,再在模型中建立隐形词汇和该词汇语义之间的关系。关系建立之后,输入需要翻译的文本信息,自动标注算法先对文本信息进行读取,判断其是否可能有隐形词汇,若该文本不含隐形词汇,则直接进行翻译。若可能含有隐形词汇,则对翻译文本中的隐形词汇进行预测,并将预测的隐形词汇进行标注。将标注的隐形词汇作为关键词,将关键词的格式进行统一处理。再将其输入到由语言模型建立的文本翻译数据库中,在该数据库中查询关键词可能的语义,判断查询到的语义和关键词之间的关系,判断其是否适用于当下的语境,若适合则留下该语义,若不适合则删除。再对关键词的不同语义的出现概率进行计算,并按照概率的高低进行排序,再将关键词和该关键词的语义按照概率高低进行降序输出。最后输出文本信息的全部翻译内容。在此过程中自动标注算法对输入的对隐形词汇的预测计算方式如式(3)所示。

$$p(t(d_n)=p(t\mid d_n)p(t)/p(d)\propto p(d_n\mid t)p(t) \tag{3}$$

式中,d_n 表示翻译文本信息,t 为关键词即隐形词汇。$p(t\mid d_n)$ 表示在给定翻译文本 d_n 的条件下,关键词 t 出现的概率,竖线表示“在某一条件下”;$\propto$ 为比例符号,是为了简化计算,忽略无关常数。

将预测的隐形词汇进行标记后,需要找到隐形词汇和语音之间的关系,计算方式如式(4)所示。

$$p(w_i\mid t)=\frac{N_t}{N_t+\mu}\cdot\frac{f(w_i,t)}{N_t}+\left(1-\frac{N_t}{N_t+\mu}\right)\cdot\frac{f(w_i,C_{tr})}{N_c} \tag{4}$$

式中,N_t 表示隐形词汇 t 在输入文本中出现的次数,$f(w_i,t)$ 为隐形词汇语义 w_i 在隐形词汇标注过的翻译文本中出现的次数和,$f(w_i,C_{tr})$ 代表隐形词汇语义 w_i 在训练文本集合中出现的次数和,μ 表示平滑因子,N_C 为整个训练文本的集合。将隐含词汇和该词汇语义的关系找到后,对隐形词汇可能语义的出现概率进行计算,计算方式如式(5)所示。

$$p=\frac{f(t,C_{tr})}{\sum_{t\in T}f(t,C_{tr})} \tag{5}$$

综合上述计算结果,由式(6)可得翻译文本中隐形词汇的语义结果列表。

$$R(t)=\operatorname{argmax} X_t^n\in{}_M p(t\mid d_n) \tag{6}$$

式中,M 表示翻译文本中所有关键词即隐形词汇的集合。

2 算法应用效果分析

为验证基于 Transformer 语言模型的自动标注算法的优越性,此次研究对基于 Transformer 语言模型的自动标注算法(算法 1),基于深度学习图像的自动标注算法(算法 2)、基于集成分类的自动标注算法(算法 3)、基于隐含狄利克雷分布(latent dirichlet allocation, LDA)语言模型的自动标注算法(算法 4)进行了模拟实验,对实验结果进行分析,4 种算法的标注准确率和误差率如图 3 所示。

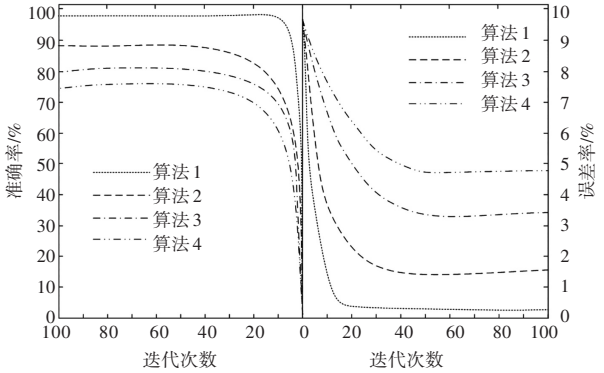
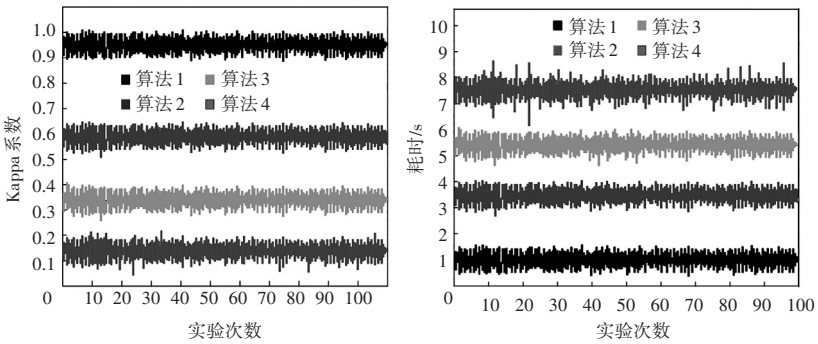


图 3 4 种算法的准确率和误差率对比

Fig. 3 Comparison of accuracy and error rate among the four algorithms

由图 3 可得,4 种算法的准确率随着迭代次数的增加而增加,而误差率随着迭代次数的增加而降低。由图 3 可得,算法 1 在迭代次数大于 15 后,该算法的准确率基本稳定下来,准确率为 98.5%;而算法 2 在迭代次数大于 20 后,准确率才达到最高,准确率最高为 88.7%,算法 3 和算法 4 都在迭代次数达到 30 后,准确率才达到最大,最大准确率分别为 80.2%和 75.4%。而算法 1、算法 2、算法 3、算法 4 的误差率分别为 0.2%、1.5%、3.4%和 4.7%。由该实验结果可知,基于 Transformer 语言模型的自动标注算法的准确率最高,误差率最低。之后再对 4 种算法的计算时间和算法 Kappa 系数进行对比,对比结果如图 4 所示。

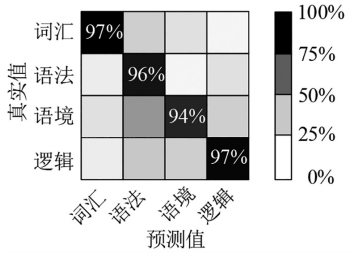


(a) 4 种算法 Kappa 值 (b) 4 种算法计算速度

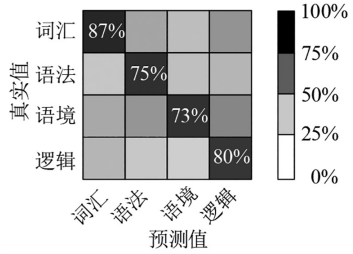
图 4 4 种算法的 Kappa 系数和计算时间对比

Fig. 4 Comparison of Kappa coefficient and calculation time among the four algorithms

Kappa 系数表示算法的一致性检验,用于衡量算法的分类精度,Kappa 系数的计算结果在 0~1 之间。一般分为 5 组:数值在 0~0.2 之间表示算法的一致性极低,0.2~0.4 表示算法的一致性一般,在 0.4~0.6 之间表示算法的一致性较好,在 0.6~0.8 之间表示算法具有高度的一致性,在 0.8~1.0 之间表示该算法的预测结果与实际结果几乎完全一致。由图 4(a)可得,算法 1 的 Kappa 系数为 0.93,即算法 1 的预测值与实际值几乎完全相同,算法 2 和算法 3 的 Kappa 系数的值分别为 0.59 和 0.31 表示这两种算法的一致性较好,算法 4 的 Kappa 系数为 0.15,则算法 1 的一致性极低。由图 4(b)可得,算法 1 的计算耗时是 4 种算法中耗时最短的,平均耗时为 1.1 s,而算法 2、算法 3、算法 4 计算平均耗时分别为 3.3 s、5.4 s 和 7.8 s。并且算法 1 的耗时稳定,波动幅度最低,而算法 4 的耗时极其不稳定。所以由该实验结果可得,基于 Transformer 语言模型的自动标注算法的预测结果与实际结果的一致性最高、计算速度最快,耗时最短,稳定性最强综合性能显著优于其他对比算法。所以此次研究采用基于 Transformer 语言模型的自动标注算法。再对使用该算法进行优化的翻译软件翻译效果进行分析,对比整体文本翻译的词汇准确率、语法准确率、语境准确率、逻辑准确率与传统的翻译软件进行对比,结果如图 5 所示。



(a) 优化后的翻译软件的翻译准确率



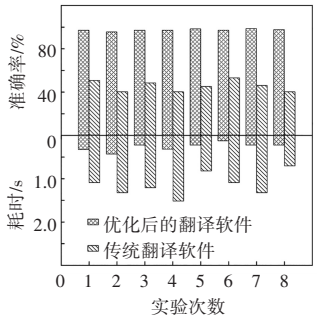
(b) 传统翻译软件的翻译准确率

图 5 翻译准确率指标对比

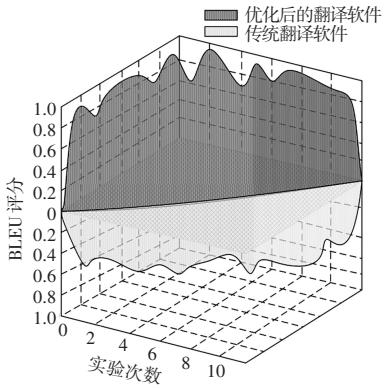
Fig. 5 Comparison of translation accuracy indicators

图 5 表示用混淆矩阵两种翻译软件的翻译准确率结果,混淆矩阵中主对角线上的百分比表示翻译正确的各项指标的百分比,左下三角的元素表示漏翻译的文本信息指标的比例,右上三角的元素表示翻译错误的文本信息指标的比例。由图 5(a)可得经过自动标注算法优化后的翻译软件中翻译文本的词汇翻译准确率、语法翻译准确率、语境翻译准确率和逻辑翻译准确率分别为 97%、96%、94%、97%,各个指标的准确率都比较高。由图 5(b)可得,传统翻译软件对翻译文本的翻译准确率远低于优化后的翻译软件。传统翻译软件中翻译文本的词汇翻译准确率、语法翻译准确率、语境翻译准确率和逻辑翻译准确率分别为 87%、75%、73%和 80%。由该实验可得,自动标注算法对语境翻译的准确率提高得最多,提高了 19%。再对使用算法优化后的翻译软件对隐形词汇语义的翻译准确率、整体翻译耗时以及双语替换测评(bilingual evaluation understudy, BLEU)评分进行对比,结果如图 6 所示。

由图 6(a)可得,经自动标注算法优化后的翻译软件对隐形词汇语义的翻译准确率提高到了 98.9%;并且优化后的翻译软件整体翻译平均耗时为 0.48 s 远低于传统模型的 1.4 s。BLEU 评分表示机器翻译结果与参考翻译结果的相似度,BLEU 评分的范围为 0~1 之间,该评分越高,表示翻译结果越相似。由图 6(b)可得,经自动标注算法优化后的翻译软件的 BLEU 评分平均为 0.92 分,而传统翻译软件的 BLEU 评分仅有 0.73 分。由上述实验可得,使用基于 Transformer 语言模型的自动标注算法对翻译软件进行优化,可以提高整体文本翻译和隐形词汇语义翻译的准确率。并且降低翻译时间。



(a) 隐形词汇语义翻译准确率和整体翻译耗时



(b) BLEU 评分

图 6 隐形词汇翻译准确率和 BLEU 评分对比

Fig. 6 Comparison of implicit word translation accuracy and BLEU score

3 结论

针对目前文本翻译时,翻译文本的隐形词汇难以被正确翻译的问题,此次研究将 Transformer 语言模型和自动标注算法进行融合,提出了一个基于 Transformer 语言模型的自动标注算法。对该算法(算法 1)、基于深度学习图

(下转第 139 页)