基于对比学习和排名一致性的古代汉语翻译质量评估模型

李怀明　邵艳秋　李炜*
北京语言大学，信息科学学院，
国家语言资源监测与研究平面媒体中心，
北京市海淀区学院路15号，100083
18811607325@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

摘要

当前，虽然机器翻译的自动评估技术已展现出良好的性能，但将它们应用于古代汉语到现代汉语的翻译场景时效果并不理想。一方面，这些传统方法能较好地比较质量差异较大的译文的好坏，但是在评估质量相差不大的译文时往往难以区分优劣。另一方面，古代汉语的省略和复杂句式常导致翻译过程中出现漏译现象，而传统评估指标往往会给这类较差的译文偏高的分数。在本文中，我们提出了一种基于对比学习和排名一致性的古代汉语到现代汉语的翻译质量评估模型（**CRATE**）。该模型通过确保语义相似度和匹配度的排名一致性捕捉译文质量的细粒度排名信息。另外，我们在使用对比学习方法训练译文跟原文的匹配模型时，将原文自身作为负样本，有效解决了传统评估指标在译文出现漏译情况下仍给出高评分的问题。为了证明我们模型的有效性，我们构建了高质量的古代汉语到现代汉语翻译的人工评分测试集。实验结果表明，我们的模型优于强大的基线，与人类评分取得了更显著的相关性。

关键词： 翻译评估；古代汉语；细粒度排名；对比学习

Ancient Chinese translation quality evaluation model based on contrastive learning and ranking consistency

Huaiming Li　Yanqiu Shao　Wei Li*
Information Science School, Beijing Language and Culture University,
Language Resources Monitoring and Research Center,
15 Xueyuan Road, HaiDian District, Beijing, 100083
18811607325@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

Abstract

Currently, although automatic evaluation techniques for machine translation have shown good performance, their results are not ideal when applied to translation scenarios from ancient Chinese to modern Chinese. On the one hand, while these traditional methods can better compare translations with large differences in quality, they are often difficult to distinguish between translations with similar quality. On the other hand, omissions and complex sentence patterns in ancient Chinese often lead to missed translations during the translation process, and traditional evaluation metrics tend to give higher scores to such poor translations. In this paper, we propose a translation quality evaluation model from ancient Chinese to modern Chinese (**CRATE**) based on contrastive learning and ranking consistency. This model captures fine-grained ranking information of translation quality by ensuring ranking consistency for semantic similarity and matching. In addition, when using the contrastive learning method to train

* 通讯作者　Corresponding Author

第二十三届中国计算语言学大会论文集，第171页–第182页，太原，中国，2024年7月25日至28日。
卷1：主会论文

the matching model between the translation and the source text, the source text itself is used as a negative sample, which effectively solves the problem of traditional evaluation metrics still giving high scores when the translation is missing. To demonstrate the effectiveness of our model, we constructed high-quality human annotation test sets of ancient Chinese to modern Chinese translations. Experimental results show that our model outperforms strong baselines and achieves more significant correlations with human ratings.

Keywords: Translation evaluation , Ancient Chinese , Fine-grained ranking , Contrastive learning

1 引言

古代汉语作为一门历史深厚的语言体系，不仅承载了中国丰富的历史遗产和哲学思想，更是理解中华文明的核心基石。然而，随着时间的推移，语言的演变使古代汉语与现代汉语之间形成了显著的差异，从而对阅读和理解构成了重大的障碍。因此，将古代汉语精确地翻译成现代汉语显得尤为关键。鉴于这一挑战，开发一种科学且客观的古代汉语翻译质量评估方法变得尤为迫切。一个准确的评估方法不仅可以增强翻译的精确性和可靠性，而且能确保中国传统文化的精髓被准确传递。总之，古代汉语的翻译质量评估不仅对于语言学研究具有重要意义，同时对于文化传承和国际文化交流亦具有不可估量的价值。

早期的机器翻译主要依赖于人工评价。Church and Hovy (1993)从流利度、忠实度和理解力三个维度对翻译进行评估。Lommel et al. (2014)提出了多维度质量指标（MQM），详细列出了机器翻译容易出现的八种错误类型。这些评估方法至今仍对翻译评估产生影响。然而，人工评估不仅成本高，还易受评估人员主观性的影响。随着技术进步，机器翻译的自动评估方法应运而生。这些自动评估方法主要分为基于字符串匹配和基于语言模型两大类。基于字符串匹配的方法涵盖了编辑距离 (Su et al., 1992)、准确率 (Papineni et al., 2002)）和召回率 (Lin, 2004)等评估指标，因其操作简便，在机器翻译评估领域得到了广泛的应用。然而，这些指标主要依赖于字面上的词汇匹配，未能充分挖掘句子的深层语义，忽视了语言的多样性与复杂性。相反，基于语言模型的评估方法 (Zhang et al., 2020)能够通过理解句子的深层语义来有效处理同义的问题，在评价质量差异显著的译文时表现出色。但是，在区分质量相近的译文时，这类评估方法的效果往往不够理想。

此外，古代汉语是一种高度精练且紧凑的语言，常涉及省略等复杂语法结构，这使得将其翻译成现代汉语时容易出现漏译现象。在图1中，我们展示了几个翻译过程中出现漏译的示例。样例1中的“**妻子**”在古代汉语中表示“**妻子和孩子**”的含义，而现代汉语中通常仅指代“**妻子**”。样例2中的“**咸秩群祀**”和样例3中的“**考系廷尉**”过于简练，导致翻译模型难以准确捕捉其含义。这些因素均会导致漏译现象。然而，传统评估指标主要关注语义相似度，常对这类质量较低的译文给出偏高的分数。鉴于现有机器翻译评估指标在评价古代汉语翻译质量方面存在不足，本文旨在提出一种更精确的评估方法，以促进古代文本翻译研究的发展。

在本文中，我们提出了一个基于对比学习和排序一致性算法的古代汉语到现代汉语的翻译质量评估模型。该模型的基本思想是，当古代汉语有参考译文以及多个机翻译文时，这些机翻译文在与参考译文的语义相似度方面上的排名以及机翻译文在与古代汉语的翻译对匹配方面上的排名之间，应该表现出高度的一致性。我们通过确保机翻译文在语义相似度和匹配度两方面的**排名一致性**，来捕捉译文质量的**细粒度排名信息**。另外，我们在采用对比学习训练原文和译文的匹配度模型时，通过**使用原文自身作为负样本**，有效解决了传统评估方法对存在**漏译情况**的机器翻译给予过高评分的问题。此外，为了让对比学习在小批量数据中学习到更有效的信息，我们通过多种破坏并重构译文的方法构建了难负样本。

为了证明我们方法的有效性，我们构建了高质量的古代汉语翻译人工评估结果的测试集。实验结果表明，我们的模型性能超过了现有的主流翻译质量评估指标，与人类评分取得了更显著的相关性。此外，通过分析质量接近的译文与古代汉语之间距离的方差，证明了我们的方法

样例1	**古代汉语：** 浑清素在公，妻子不免於饥寒。 **参考译文：** 郑浑为官的时候清正廉洁，家中妻子孩子经常忍饥挨饿。 **机翻译文：** 浑清素在公，妻子不免于饥寒。
样例2	**古代汉语：** 六月庚子，初祀五岳四渎，咸秩群祀。 **参考译文：** 六月一日，前往祭祀五岳名山，百官都有秩序地参加祭礼。 **机翻译文：** 六月庚子日，最初的祭祀五岳四渎，咸秩群祀。
样例3	**古代汉语：** 南阳圭泰尝以言迕指，考系廷尉。 **参考译文：** 南阳人圭泰曾在言语上违背了他们的指示，就被廷尉查考抓捕。 **机翻译文：** 南阳人圭泰曾用言语冒犯指，考系廷尉。

Figure 1: **古代汉语翻译过程中出现漏译情况的样例。**机翻译文来源于百度翻译[1]，漏译位置用不同颜色字体进行了标记。

能够提供更准确的细粒度排名信息。同时，我们单独对存在漏译情况的译文进行了实验分析，结果验证了我们在对比学习中将原文作为负样本的策略的有效性。

本文的主要贡献有：

- 我们提出了一个古代汉语翻译质量评估模型（CRATE），其通过确保机翻译文在语义相似度和匹配度两方面的排名一致性来捕捉细粒度的译文质量排名信息，改善了传统方法难以辨别质量相近译文优劣的问题。
- 在通过对比学习训练原文和译文的匹配模型时，我们将原文自身用作负样本，有效解决了传统评估指标在评价存在漏译现象的译文时倾向于给出过高分数的问题。
- 我们构建了高质量的古代汉语到现代汉语翻译的人工评估结果的数据集。实验结果表明，我们的模型的评估结果与人工注释结果高度相关，超过了当前的评估指标。

2 相关工作

机器翻译的自动评估方法最初基于字符串匹配，涵盖了编辑距离、准确率和召回率等方面。编辑距离的评估方法主要包括词误差率（WER）、位置无关错误率（PER）及翻译错误率（TER）。其中，WER (Su et al., 1992)通过计算机译文与参考译文之间的编辑距离来评估翻译质量，但对词序变化的惩罚过重，处理效果不佳。为了解决这一问题，PER (Tillmann et al., 1997)采用忽略词序的方法，而TER (Snover et al., 2006)则提出将词语移动的惩罚减至最低。基于准确率的BLEU (Papineni et al., 2002)和基于召回率的ROUGE (Lin, 2004)都是通过计算机译文与参考译文的n-gram重叠来进行评估。然而，这些指标主要关注词汇层面的重叠，未能充分体现句子的语义，对同义词或相似表达的处理尚显不足。为了弥补这一缺陷，METEOR (Banerjee and Lavie, 2005)引入了同义词词表，但其过分依赖于WordNet数据库，在中文翻译评估中的适用性有限。

随着预训练语言模型技术的快速发展，评估指标开始采用这些模型来理解句子的深层语义。例如，BERTscore (Zhang et al., 2020)利用语言模型来计算机翻译文跟参考译文之间的语义相似度，同时引入逆文本频率指数（IDF）来考虑词的重要程度。尽管这类指标在区分出质量相差较大的译文时表现出色，但在评估质量接近的译文方面效果不佳。当前，专门用于古代汉语到现代汉语翻译的质量评估方法相当有限。在这一领域，常用的评估方法仍然是基于字符串匹配的传统方法，如 (Li et al., 2018; Liu et al., 2020; Zhou and Si, 2023)在评估古代汉语翻译质量时仍然采用传统的BLEU评估指标。为了更好地满足古代汉语翻译的特殊需求，Yang et al. (2021)提出了针对古汉语翻译的自动评估指标DTE，其结合了BLEU和句子相似度评价方法的优点。然而，当机翻译文中出现漏译现象时，这些基于语义相似度的评估指标仍然倾向于给出偏高的评分。

[1] https://fanyi.baidu.com/

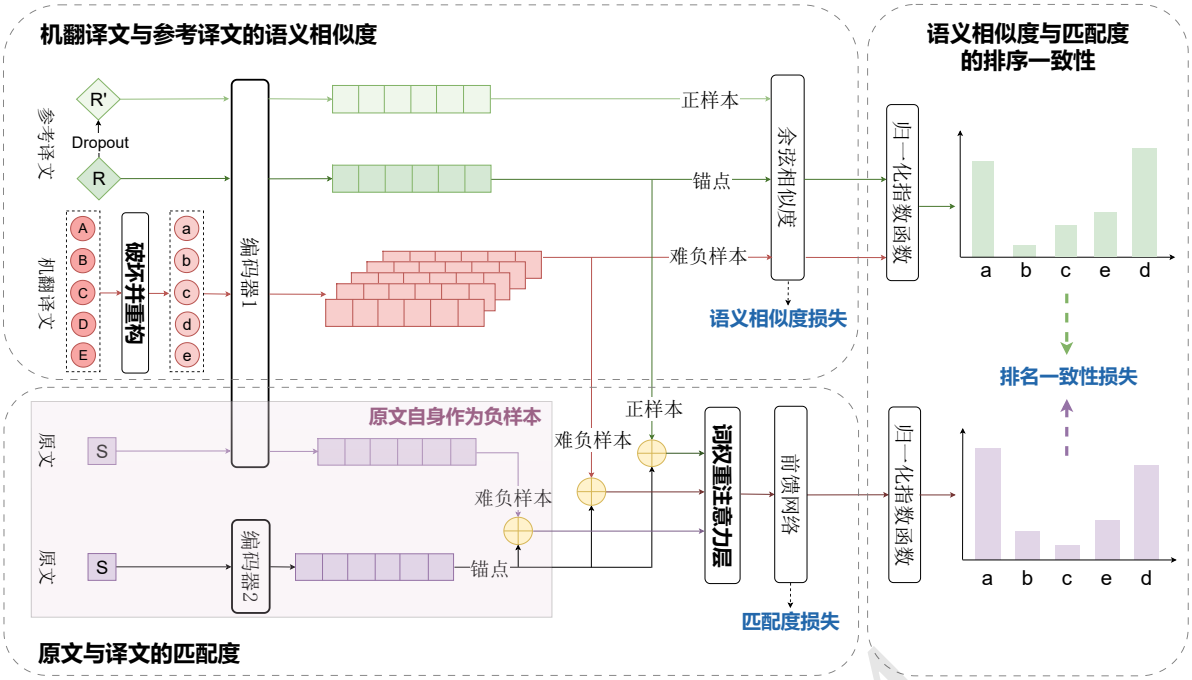

Figure 2: **古代汉语翻译评估模型CRATE的总体架构。** CRATE通过排名一致性来捕获细粒度的译文质量排名，并在对比学习中将原文作为负样本来解决漏译问题。该模型的总体训练目标是：(1)机翻译文和参考翻译之间语义相似度的对比学习损失；(2)机翻译文与原文的匹配方度的对比学习损失；(3)语义相似度与匹配度之间的排序一致性损失。

3 基于对比学习和排序一致性的古代汉语翻译质量评估模型

在本节中，我们介绍古代汉语翻译质量评估模型（CRATE）。CRATE的基本思想是通过确保语义相似度和匹配度两方面的排名的一致性来获得准确的细粒度的译文质量排名信息。具体而言，我们分别获取机翻译文在与参考译文的语义相似度方面上的排名以及机翻译文在与原文的匹配度方面上的排名，并确保它们之间的一致性。如图2所示，我们的方法由三部分组成：(1)机翻译文与参考译文的语义相似度（3.1节）；(2)原文与译文的匹配度（3.2节）；(3)语义相似度与匹配度的排名一致性（3.3节）。我们在3.4节介绍推理过程中译文质量分数的计算方法。

3.1 机翻译文与参考译文的语义相似度

该部分旨在计算机翻译文与参考译文之间的语义相似度分数。具体而言，我们通过对比学习的方法拉近语义相似的句子之间的距离，而让不同语义的句子距离更远，以此让编码器有效学习句子的向量表示。我们使用参考译文作为锚点。参考前人的工作 (Gao et al., 2021)，我们对参考译文应用dropout方法来生成语义相似的正样本。同时，我们采取调用语言模型API或者直接将古代汉语输入到翻译软件中，以获取多个机翻译文，之后再对它们进行破坏并重构的方法来构造难负样本。

难负样本生成 受到BART (Lewis et al., 2020)模型预训练过程中引入噪声的方式的启发，我们在构造负样本时通过不同的噪声随机破坏机翻译文，包括掩码、删除词语、随机词语替换、随机词语插入、置换词序等，再使用预训练语言模型BERT (Devlin et al., 2019)进行重构。该模型使用破坏并重构任务进行预训练，以便更好地重建被破坏的机翻译文。通过这种方法生成的是负样本更难，能够使模型在小批量数据中学习到更加有效的信息，并且可以更好地模拟翻译过程中可能遇到的各种错误。我们采用适当的比例来破坏文本，因为添加噪声太多会导致负样本没有包含有效的信息，添加噪声太少会导致负样本跟正样本过于接近，这样都不利于模型训练。我们在4.1节介绍了具体的细节。

之后，我们将这些处理好的锚点、正样本和负样本输入到预训练语言模型中，获得它们的句子向量表示，并通过计算余弦相似度来评估它们之间的语义相似性。在模型训练过程中，我

们以InfoNCE损失函数 (van den Oord et al., 2018)作为训练目标，因为其通过最大化正样本对的相对概率并最小化负样本对的相对概率，强化了模型对特征差异的敏感度。如式1所示：

$$\mathcal{L}_{similarity} = -log\frac{e^{sim(f_1(x_i),f_1(\overline{x}_i))/\tau_1}}{\sum_{j=1}^{N} e^{sim(f_1(x_i),f_1(x_j))/\tau_1}} \tag{1}$$

其中，$f_1(\cdot)$是通过预训练模型得到的句子向量表示，x_i 是参考译文（锚点），$\overline{x}_i$ 是参考译文经过dropout后的文本（正样本），x_j 是机翻译文经过破坏并重构后的文本（负样本），N是负样本的数量，τ_1是温度超参数，$sim(f(x_i), f(x_j)) = \frac{f(x_i)^\top f(x_j)}{\|f(x_i)\| \cdot \|f(x_j)\|}$ 是余弦相似度。

3.2 原文与译文的匹配度

该部分旨在计算机翻译文与古代汉语之间的匹配分数。我们同样采用对比学习的方法进行训练，其中原文作为锚点，参考译文作为正样本，而破坏并重建的机翻译文作为负样本。另外，鉴于传统评估指标倾向于给那些存在漏译情况的译文较高的分数，我们还将原文自身作为了负样本，这样可以有效地降低这类质量较差的译文跟原文之间的匹配分数。

我们使用3.1节的预训练模型f_1来获取参考译文x_i的句子表示h_i（正样本）、经过破坏并重构之后的机翻译文x_j的句子表示h_j（负样本）以及原文x_s的句子表示h_s（负样本）。此外，我们使用guwenbert [2]作为预训练语言模型f_2来获得原文x_s的句子表示$h_{s'}$（锚点）。随后，我们分别将h_i、h_j、h_s与$h_{s'}$进行拼接。这个过程如式2所示：

$$h_{t+s'} = concat(h_t, h_{s'}),\ t\epsilon\{i, j, s\} \tag{2}$$

词权重注意力层 Zhang et al. (2020)认为在评估译文的翻译质量时，词的重要程度并不相同，于是其使用IDF来计算不同词的权重信息。在本文中，我们应用注意力机制 (Vaswani et al., 2017)让模型集中处理更重要的语义信息。具体而言，我们使用权重向量q沿维度抽取拼接后的特征表示, 如式3所示：

$$\mathcal{V}_{t+s'} = q \cdot h_{t+s'},\ t\epsilon\{i, j, s\} \tag{3}$$

其中，q的值服从概率分布，且每行的概率的总和都为1。

随后，我们将抽取完的特征向量输入到前馈神经网络层，分别预测得到它们之间的匹配分数，如式4所示。

$$score_{t+s'} = \omega_1 ReLU(\omega_2 \mathcal{V}_{t+s'} + b_1) + b_2,\ t\epsilon\{i, j, s\} \tag{4}$$

其中，ω_1、ω_2、b_1 和b_2 是模型参数。

最后，我们通过对比学习最大化原文与参考译文之间的匹配分数，同时最小化原文与低质量译文之间的匹配分数。训练目标如图5所示：

$$\mathcal{L}_{match} = -log\frac{e^{score_{i+s'}/\tau_2}}{e^{score_{s+s'}/\tau_2} + \sum_{j=1}^{N} e^{score_{j+s'}/\tau_2}} \tag{5}$$

其中，τ_2 是温度超参数，N是负样本的数量。

3.3 语义相似度与匹配度的排名一致性

通过对比学习，我们能够有效地区分正样本跟负样本这种语义差距较大的样本。然而，模型并没有学习到负样本之间细粒度的差异，其仍然无法区分语义接近的样本。因此，我们通过确保语义相似度和翻译对匹配度两方面的排名的一致性来捕获译文的细粒度排名信息。具体而言，我们利用3.1节的语义相似度模型计算多个机翻译文跟参考译文的语义相似度的排名列表，利用3.2节的匹配模型计算多个机翻译文跟原文的匹配度的排名列表。

参考 Liu et al. (2023)的做法，我们利用ListNet (Cao et al., 2007)将排序列表看做概率分布。由于KL散度（Kullback-Leibler Divergence）不是对称的，所以我们通过最小化两个顶部概率分布（top one probability distributions）之间的JS散度（Jensen-Shannon divergence）来确保语义相似度和匹配度两方面的排名一致性，如式6所示。

[2] https://github.com/Ethan-yt/guwenbert

$$\begin{aligned}\mathcal{L}_{consistency} &= \sum_{i=1}^{N} JS(P_i \| Q_i) \\ &= \frac{1}{2}\sum_{i=1}^{N} KL(P_i \| \frac{P_i+Q_i}{2}) + \frac{1}{2}\sum_{i=1}^{N} KL(Q_i \| \frac{P_i+Q_i}{2}) \\ &= \frac{1}{2}\sum_{i=1}^{N} (P_i\ log(\frac{2P_i}{P_i+Q_i}) + Q_i\ log(\frac{2Q_i}{P_i+Q_i}))\end{aligned} \tag{6}$$

其中P_i和Q_i分别表示语义相似度列表和匹配分数列表的概率分布。

简而言之，$L_{similarity}$和L_{match}通过对比学习区分正负样本，忽略了负样本之间的差异性。$L_{consistency}$通过确保语义相似度以及匹配度两方面的排名一致性来捕获负样本之间细粒度的译文质量排名信息。我们模型的总的损失函数为：

$$\mathcal{L}_{local} = \mathcal{L}_{similarity} + \alpha\mathcal{L}_{match} + \beta\mathcal{L}_{consistency} \tag{7}$$

其中α 和β 是超参数。

3.4 推理

在推理过程中，如果评估时无参考译文，我们需将机翻译文输入到编码器f_1中获取句子表示h_j，并将古代汉语输入到编码器f_2中获得句子表示$h_{s'}$。然后，将这两个句子表示进行拼接，并输入到注意力层和前馈神经网络中，计算出它们的匹配分数$score_{j+s'}$。

如果评估时有参考译文，我们则将参考译文和机翻译文输入到编码器f_1中获得h_i和h_j，并计算它们之间的余弦相似度。同样地，我们也需要计算机翻译文与古代汉语之间的匹配分数$score_{j+s'}$。最终的评估分数通过综合这两个分数来获得，公式如下：

$$score_{local} = \lambda score_{j+s'} + \mu sim(h_i, h_j) \tag{8}$$

其中λ 和μ 是超参数。

4 实验部分

4.1 设置

参考 Ranasinghe et al. (2020)的工作，我们利用[CLS]标记对应的嵌入作为输入句子的表示。编码器在模型训练过程中更新参数。我们采用AdamW (Loshchilov and Hutter, 2019)作为优化器，并设置学习率为2e-5。此外，超参数α和β的值设置为1，λ 和μ的值设置为0.5。我们设置负样本的数量为25，且调整温度超参数τ_1 和τ_2为0.3（详情见4.3.4节）。另外，破坏机翻译文的比例我们设定为0.5。模型经过5个训练周期后，我们保留表现最佳的检查点，以便对测试集进行最终评估。

数据 本文的数据抽取自东北大学提供的一个包含古代汉语与现代汉语翻译对齐的数据集 [3]。我们随机抽取了20,550条数据对，其中20,000条数据对被划分为训练集。我们使用多个翻译模型得到古代汉语的机翻译文，包括gpt-3.5-turbo [4]，GLM-3-Turbo [5]，文心一言 [6]，wenyanwen-ancient-translate-to-modern [7]和百度翻译 [8]。对于模型，我们通过API调用方式获得译文；对于百度翻译，则直接输入古代汉语。余下的550条数据对经过筛选，剔除不合适的数据后，选取500条数据对中的古代汉语输入到wenyanwen-ancient-translate-to-modern来获取机翻译文。随后，我们根据原文和参考译文对这500条机翻译文进行人工评分，并将其作为测试集（UTS）。

[3] https://github.com/NiuTrans/Classical-Modern
[4] https://platform.openai.com/docs/models/gpt-3-5-turbo
[5] https://open.bigmodel.cn/usercenter/apikeys
[6] https://console.bce.baidu.com/qianfan/ais/console/applicationConsole/application
[7] https://huggingface.co/raynardj/wenyanwen-ancient-translate-to-modern
[8] https://fanyi.baidu.com/

评估指标 受到 Huang et al. (2023)工作的启发，我们通过计算模型预测与人工评分的皮尔逊相关系数(Pearson Correlation Coefficient)和斯皮尔曼等级相关系数(Spearman's Rank Correlation Coefficient)来评估模型性能。皮尔逊相关系数是关于两个随机变量之间的线性关系的统计度量，使用数据样本值本身进行计算，计算公式如式9所示。斯皮尔曼等级相关系数是随机变量之间单调关系的统计度量，使用数据样本排位位次值进行计算，计算公式如式10所示。

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{9}$$

其中，$\overline{x}$和$\overline{y}$是变量的均值。

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}((R(x_i) - \overline{R(x)}) \cdot (R(y_i) - \overline{R(y)}))}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})^2) \cdot (\frac{1}{n}\sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2)}} \tag{10}$$

其中，$R(x_i)$和$R(y_i)$是变量x和y的位次，而$\overline{R(x)}$和$\overline{R(y)}$是对应排名的平均位次。

基线 我们将CRATE与其他评估指标进行了对比，包括BLEU、ROUGE和BERTscore。此外，我们还与古代汉语翻译质量评估指标DTE进行了比较。对于基于预训练模型的评估指标，我们尝试在$\text{BERT}_{\text{base}}$和$\text{RoBERTa}_{\text{base}}$ [9]上进行实验，我们的方法只更换f_1。另外，我们还在SimCSE[10]上进行了实验，通过直接计算机翻译文跟参考译文的语义相似度来评估译文质量。

4.2 实验结果

如表1所示，我们的方法在PCC和SRCC上与人类标注的相关性都是最高的，这验证了我们方法的有效性。在两个评估方法上，基于字符串的指标要明显差于基于预训练模型的指标，这证明前者确实只注意句子的表层词汇而忽视了深层语义分析。SimCSE的性能比BERTScore好很多，因为它通过对比学习获得了更准确的句子表示。此外，DTE的性能比SimCSE略好，因为其结合了BLEU以及句子相似度的优势，更贴合古文翻译的特点。值得注意的是，我们的方法无论基于BERT还是基于RoBERTa，在两个指标上都优于其他的基线。

预训练模型	方法	PCC	SRCC
Non-BERT	BLEU	0.4903	0.5087
	ROUGE	0.5125	0.5032
$\text{BERT}_{\text{base}}$	+BERTScore	0.5971	0.6013
	+SimCSE	0.6530	0.6769
	+DTE	0.6927	0.6834
	+CRATE (Ours)	**0.7153**	**0.7297**
$\text{RoBERTa}_{\text{base}}$	+BERTScore	0.6264	0.6427
	+SimCSE	0.6752	0.6618
	+DTE	0.6995	0.7019
	+CRATE (Ours)	**0.7328**	**0.7285**

Table 1: **CRATE和基线在测试集上的实验结果。** 计算我们的模型与基线预测跟人工评分之间的皮尔逊相关系数（PCC）和斯皮尔曼等级相关系数（SRCC）。结果表明我们的方法无论基于BERT还是基于RoBERTa，在两个指标上都与人工评分具有最高的相关性。

4.3 分析与讨论

4.3.1 消融实验

在本节，我们通过消融实验探讨我们提出的方法中各个组成部分的作用。结果如表2所示。首先，当我们移除排名一致性模块时，模型无法获得精确的译文质量排名信息，导致在两个评

[9] https://github.com/ymcui/Chinese-BERT-wwm
[10] https://huggingface.co/hellonlp/simcse-roberta-base-zh

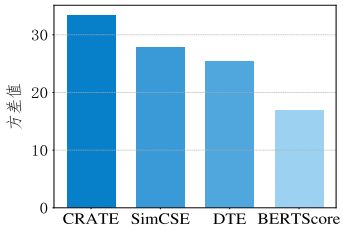

Figure 3: **锚点与负样本之间欧氏距离的方差值。**我们提出的方法计算得到的方差值最大，证明其确实可以捕获负样本之间细微的差别。

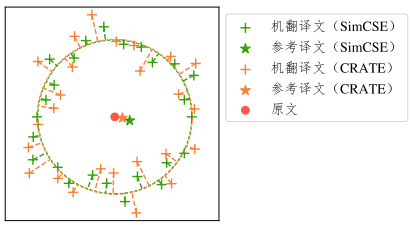

Figure 4: **锚点与负样本之间欧氏距离的案例可视化。**圆圈半径为平均欧式距离，通过样本点到圆周的垂直线长度可以看出，CRATE的样本分布比SimCSE更分散。

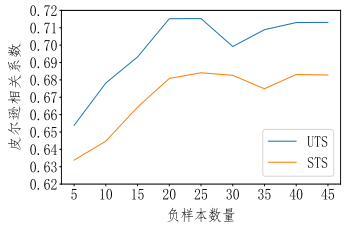

Figure 5: **探究负样本数量对于模型性能的影响。**我们对UTS和STS测试集在不同负样本数量下进行了模型预测与人工评分之间的皮尔逊相关系数计算。

估指标上的结果有所下降。其次，移除对比学习后，模型失去了区分正负样本之间差异的能力，从而导致性能显著降低。另外，当我们仅使用匹配模型或语义相似度模型都会导致性能下降，两者结合使用时效果最佳。再者，当我们移除破坏并重构模块时，性能显著下降。这是因为机翻译文跟参考译文之间的质量差距通常不大，使用它们作为对比学习的正负样本时，模型可能学习到不准确的信息。最后，移除词权重注意力层后，性能略有下降，这表明它在我们的建模中的有效性。值得注意的是，当我们训练匹配模型时不使用原文作为负样本，结果并没有什么变化，这是因为测试集中的译文大多翻译完全，并没有出现漏译的情况，我们在4.3.3节进行了详细的分析。

方法	PCC	SRCC
CRATE	0.7153	0.7297
w/o 排名一致性	0.6924	0.6985
w/o 对比学习	0.6379	0.6149
w/o 匹配度&排名一致性	0.6785	0.6773
w/o 相似度&排名一致性	0.6833	0.6917
w/o 原文作为负样本	0.7149	0.7275
w/o 破坏并重构	0.6518	0.6397
w/o 词权重注意力层	0.7096	0.7110

Table 2: **我们的模型基于$\mathbf{BERT_{base}}$的消融实验结果。**包含所有模块的模型具有最佳的性能，这证明了我们模型的每个组件的必要性和有效性。w/o 代表删除。

4.3.2 译文质量的细粒度排名分析

4.2节的实验结果表明，与其他基准模型相比，我们提出的方法与人工评分的相关性更高。为了探讨我们方法的卓越表现是否源于其区分质量接近的译文的细微差别的能力，我们计算了机翻译文（负样本）与原文（锚点）之间的欧氏距离，并通过分析其方差来进行评估。具体而言，我们对测试集的数据生成一些负样本，并计算平均方差。方差值越大，说明区分负样本细微差异的能力越强。

结果如图3所示，我们的方法（CRATE）计算得到的负样本的方差值最大，这表明我们的方法捕捉到的负样本的质量分布更为分散，证明其确实可以捕捉到质量接近的译文之间的细微差别，从而得到准确的细粒度译文质量排名信息。相比之下，SimCSE方法虽然通过对比学习有效地区分了正样本和负样本，但其在负样本间的细微差异识别上表现较为有限，这可以从其较小的负样本间方差中观察到。此外，DTE在性能上与SimCSE相近，我们认为虽然DTE并未通过对比学习拉开样本之间的距离，但是其通过结合BLEU和语义相似度的优点弥补了这方面的不足。BERTScore计算的方差值最小，这是因为它在处理细粒度质量评估方面的能力较弱，无法有效区分质量不同的负样本。

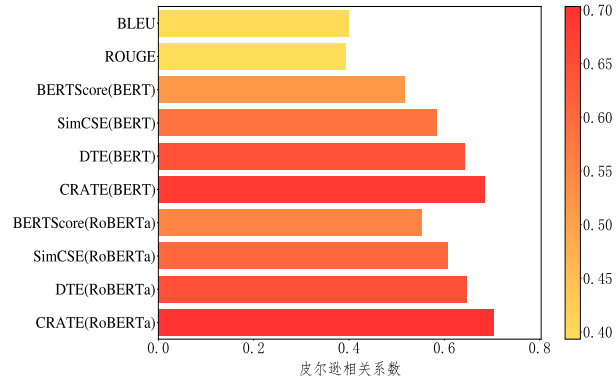

Figure 6: **CRATE和基线在存在漏译情况的测试集（STS）上的实验结果。**实验结果与UTS基本一致。基于预训练模型的评估方法明显优于基于字符串的方法，且相同的指标基于RoBERTa的性能要略优于基于BERT的性能。

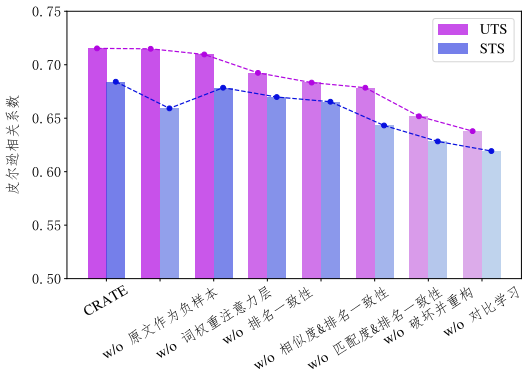

Figure 7: **CRATE在UTS和STS两个测试集上的消融实验对比。**删除“原文作为负样本”后，在几乎未存在漏译的UTS测试集上性能基本不变，在存在漏译的STS测试集上的性能明显下降。

另外，在图4中我们展示了一个具体案例。在对比分析中，我们选用SimCSE作为基准，并在图中呈现了两种方法的性能，其中圆的半径为机翻译文与原文之间的平均距离。观察SimCSE的结果，我们可以发现机翻译文的分布大多集中在圆周附近，这说明SimCSE在区分质量相近的译文方面确实存在不足。相比之下，我们的CRATE方法则呈现出更为分散的分布，表明它能有效捕捉负样本之间的细微差异。综上所述，这些发现突显了我们方法在细粒度译文质量评估方面的有效性。

4.3.3 古文漏译情况分析

为了验证模型处理存在漏译现象的译文的能力，我们另外构建了一个特殊的测试集（STS），其中的机翻译文均存在漏译现象。具体做法如下，我们从语料库中随机抽取了4,000条数据对，使用百度翻译将古代汉语翻译成现代汉语，并从中筛选出500条存在漏译的翻译。之后，我们根据原文和参考译文，对这些机翻译文进行了人工评分。结果如图6所示，在STS测试集上的结果与UTS测试集大概一致，基于预训练模型的评估指标明显优于基于字符串的性能指标，且相同指标基于RoBERTa的性能要略优于基于BERT的性能。我们在STS测试集上进行了消融实验分析（见图7），发现其性能走势与UTS测试集基本一致。值得注意的是，当移除原文作为负样本时，在UTS测试集上性能几乎不变，而在STS测试集上则有显著下降。这一结果强调了我们提出的在使用对比学习方法训练原文与译文的匹配度模型时将原文自身作为负样本的方法的有效性。

4.3.4 对比学习中超参数的影响

我们在UTS和STS两个测试集上探究了负样本数量对于模型性能的具体影响。如图5所示，在负样本数量增加的初始阶段，模型性能有明显的提升。对于UTS，当负样本数量增至20后，模型的性能提升趋缓，并且在数量增至30时性能略有下降，之后性能回升并保持相对平稳。对STS测试集的观察结果显示，模型性能在负样本数量增至25时达到较高值，增至35时略有回落，随后也表现出较为平稳的趋势。我们认为我们通过破坏并重构的方法生成了更难的负样本，所以才让我们能够用小批量的负样本获得如此的效果。这些结果表明，适量增加负样本的数量可以有效提升模型性能，但超出一定的阈值后，模型性能提升会停滞。基于这些观察，我们将负样本的数量设定为25，以达到模型性能和计算效率的最优平衡。另外，根据经验，我们将对比学习的温度系数设置为0.3。

4.4 样例分析

在本节中，我们提供了一些古代汉语翻译质量评估的具体样例，并将我们的方法与SimCSE和BERTScore两个基线进行比较。

在图8中，我们展示了对不同质量的机翻译文的细粒度评估的样例。其中，机翻译文a和b的质量较高，机器译文c、d以及e的质量较低。之前的结果（见图3）表明，SimCSE在译文质量的

古代汉语：又讨骏之时，殿下在外，实所不综。 **参考译文**：另外，声讨杨骏的时候，殿下你还在朝廷之外，委实不曾参与。	SimCSE	CRATE	Label (Rank)
机翻译文a：另外，殿下没有参与谴责杨骏，他那时候在朝廷外。	0.8115(2)	0.8346(1)	1
机翻译文b：声讨杨骏时，殿下你在朝廷外，委实是没有办法。	0.8162(1)	0.8159(2)	2
机翻译文c：我们在责备骏的时候，你在外，实在是不了解。	0.5516(4)	0.5647(3)	3
机翻译文d：在讨论杨骏的时候，你还在外面，实在是过分。	0.5405(5)	0.5328(4)	4
机翻译文e：你在夸奖杨骏的时候，确实并没有苦衷。	0.5584(3)	0.4935(5)	5

Figure 8: **不同质量的机翻译文的细粒度评估结果样例。**结果包含了评分和排名，其中的标签由人工进行标注。结果表明，SimCSE得到的细粒度排名并不准确，而我们的模型CRATE对译文质量的细粒度排名与标签完全一致。

样例	BERTScore	CRATE	Label
古代汉语：尊兄**应期赞世**，**配业光国**，魄兆见矣。 **参考译文**：尊兄您顺应天时，把握时机，辅佐光大国家事业，智慧的光芒已经显现。 **机翻译文**：哥哥**应期赞世**人，**配业光国**，魄预见了。	0.7648	0.3573	0.2
古代汉语：今以贵人为皇后，**使行丞相事左将军向朗持节授玺绶**。 **参考译文**：现在封贵人为皇后，派代理丞相事宜的左将军向郎持符节授予印信绶带。 **机翻译文**：现在用贵人为皇后，**使行丞相事左将军向朗持节授玺绶**。	0.8849	0.5409	0.6

Figure 9: **存在漏译情况的译文评估结果的两组样例。**红色字体标记了机翻译文中漏译的部分以及其在古代汉语中的相应位置，标签是人工评分。结果表明，BERTScore给存在漏译的译文生成了偏高的分数，而我们的模型评分结果与人工评分更接近。

细粒度分析方面比DTE表现更优，因此我们将SimCSE作为了基准方法。通过SimCSE的评估结果，我们观察到a和b的得分显著高于c、d和e的得分，这证实了SimCSE在区分质量差距较大的译文方面的有效性。然而，其对译文a和b之间，以及c、d和e之间的排名与实际标签并不一致，这暴露了它在获取精确细粒度排名上的局限性。相反，我们提出的模型不仅明确区分了高质量和低质量的译文，而且得到的细粒度的译文质量排名信息与标签完全一致，显著展现了其在捕捉质量接近译文间微妙差异的能力。

在图9中，我们展示了两组存在漏译情况的译文的评估样例，并用红色字体高亮显示了机翻译文中漏译的部分以及其在古代汉语中的相应位置。评估结果表明，BERTScore指标明显高于人工评分。相比之下，我们模型的预测结果与人工评分更为一致，这表明我们的方法能够为存在漏译现象的译文提供更准确的评分，进一步证实了我们提出的方法的有效性和优越性。

5 结论

在本文中，我们提出了一种基于对比学习和排名一致性的古代汉语翻译评估模型（CRATE）。通过确保机翻译文在与参考译文语义相似度上的排名以及与古代汉语匹配度上的排名之间的一致性，CRATE捕获到了细粒度的译文质量排名信息。使用对比学习训练原文和译文的匹配模型时，将原文自身作为了负样本，给存在漏译情况的译文更合理的分数。在测试集上的实验结果表明，CRATE的性能超过了现有的评估方法，与人类评分具有更高的相关性。大量的实验分析也证明了我们模型中每个组成部分的有效性和合理性。尽管预训练模型表现出较好的性能，但在古文的语义理解方面仍有提升空间。在未来的研究中，我们将探讨如何利用大语言模型进行评估。然而，目前大语言模型在古文方面的训练语料极为有限，这也是我们亟需关注的重点。

致谢

本成果受国家自然科学基金项目（62306045，61872402），北京语言大学校级项目（中央高校基本科研业务费专项资金）（18ZDJ03），北京语言大学梧桐创新平台项目（21PT04），

北京语言大学研究生创新基金项目（中央高校基本科研业务费专项资金）（24YCX085）资助。

参考文献

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 129–136. ACM.

Kenneth Ward Church and Eduard H. Hovy. 1993. Good applications for crummy machine translation. *Mach. Transl.*, 8(4):239–258.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Hui Huang, Shuangzhi Wu, Kehai Chen, Hui Di, Muyun Yang, and Tiejun Zhao. 2023. Improving translation quality estimation with bias mitigation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2175–2190. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3890–3900. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2020. Ancient-modern chinese translation with a new large training dataset. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(1):6:1–6:13.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. Rankcse: Unsupervised sentence representations learning via learning to rank. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13785–13802. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5070–5081. International Committee on Computational Linguistics.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006*, pages 223–231. Association for Machine Translation in the Americas.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, August 23-28, 1992*, pages 433–439.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas, editors, *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, pages 2667–2670. ISCA.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Kexin Yang, Dayiheng Liu, Qian Qu, Yongsheng Sang, and Jiancheng Lv. 2021. An automatic evaluation metric for ancient-modern chinese translation. *Neural Comput. Appl.*, 33(8):3855–3867.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Siqing Zhou and Shijing Si. 2023. Evaluating the capability of chatgpt on ancient chinese. *CoRR*, abs/2312.15304.