Corpus Analysis Vector: A Data-driven Performance-oriented Corpus Quality Assessment For Neural Machine Translation

Anonymous ACL submission

Abstract

Neural Machine Translation (NMT) employs neural networks to model the probability distribution of the parallel corpus, with advances in network architectures resulting in a substantial enhancement in translation quality. The quality of the parallel corpus is also a significant factor in the translation quality. Despite the broad consensus on the positive correlation between corpus quality and translation quality, existing methods for assessing corpus quality fail to address the quantitative relationship between corpus quality and translation quality. It leads to the fact that corpus quality assessment has to rely on subjective experience or black-box language models to blur the relationship, divorced from the mathematical modeling of NMT. This brings unavoidable bias and unestimated impact to the NMT system. In response to the aforementioned issues, this paper proposes the Corpus Analysis Vector (CAV), a data-driven framework that mathematically formalises corpus quality by converting text sequences into matrices under the modelling of NMT. The paper employs the CAV framework to model the probability distribution of corpus and translation quality, mathematically formalising the relationship in the context of the translation accuracy prediction task. The efficacy of CAV is validated through experimentation on multiple benchmark datasets: CAV demonstrates efficacy in translation accuracy prediction by modelling the quantitative correlation between corpus quality and translation quality. The subsequent case studies are intended to illustrate the interpretability of the CAV in terms of identifying quality-critical corpus features from a data-driven perspective. It has been demonstrated that, in addition to theoretical insights, CAV also has practical utility in guiding corpus filtering, thereby enhancing NMT systems.

1 Introduction

The Neural Machine Translation (NMT) task involves the generation of the target-side language sequence from the source-side one via a neural network. The neural network stores the probability distribution in the parameters, which are optimised to approximate the distribution of the training parallel corpus. In the preceding decade, significant advancements have been witnessed in the domain of neural networks, with a notable enhancement in the caliber of translation quality. This enhancement is largely attributed to the attention mechanisms within neural network architectures.

Conversely, corpus research continues to encounter numerous challenges. Despite the extensive recognition of a positive correlation between corpus and translation quality, there remains a paucity of research investigating the quantitative relationship between the two. Existing research has superseded that quantitative relationship with heuristic, subjective experiences, or black-box language models. The mathematical modelling of corpus quality assessment methods based on these paradigms deviates from NMT due to the absence of quantitative relationships. This lack of consistency has the potential to result in the deletion of information and the introduction of errors, as well as impeding the analysis of the relationship between corpus quality and translation quality. This, in turn, has consequences for the accuracy and flexibility of downstream applications of corpus quality assessment.

In order to address the aforementioned drawbacks, this paper proposes the Corpus Analysis Vector (CAV), which directly bridges the corpus quality and translation quality, thus achieving a dual purpose. Firstly, the CAV is data-driven, utilising a mathematical framework to convert text sequences into matrices, which is grounded in the mathematical modelling of the NMT task to preserve the essential information and maintain consistency. In this manner, CAV is responsible for conveying the requisite information for the NMT task and the original features of the corpus, including the corpus

quality. Secondly, CAV is performance-oriented, which refers to translation quality in this paper, directly bridging the relationship between corpus quality and translation quality by constructing the probability distribution of CAV and translation accuracy. In this manner, the quantitative relationship between the corpus quality and the translation quality is modelled in the probability distribution of the CAV and the translation accuracy. Specifically, the paper first provides the fundamental formulation of CAV, which contains the computation of the transformation matrix. The translation accuracy prediction task is then proposed, along with the corresponding neural network to model and approximate the probability, respectively.

The experimental results, drawn from publicly available corpora, demonstrate the impressive performance of CAV in the translation accuracy prediction task. These results indicate a strong quantitative correlation between corpus quality and translation quality. This paper further provides an in-depth analysis of the experimental results using CAV, demonstrating the interpretability and great potential of CAV as a corpus quality assessment tool.

The three principal contributions of this paper are as follows:

- **Corpus Analysis Vector (CAV)**: This paper proposes a data-driven corpus feature representation method based on task-specific mathematical modelling. This approach overcomes the limitations of heuristic-driven approaches and black-box feature extraction.

- **Translation Accuracy Prediction**: The model establishes quantitative connections between corpus quality and translation quality through the probability distribution modelling of CAV representation and translation accuracy.

- **Downstream Application**: The dual utility of CAV is demonstrated in two distinct capacities: firstly, as a quantitative corpus quality assessment tool in the context of translation error analysis, and secondly, as a data filtering guidance mechanism within the framework of neural machine translation systems.

2 Background

2.1 Corpus Quality Assessment

The two primary objects of corpus quality assessment for the NMT task are the parallel corpus and the translations. The evaluation of translation quality serves to assess the performance of the NMT system by measuring the quality of the generated translations. BLEU (Papineni et al., 2002) is a metric of considerable popularity, the function of which is to calculate the precision of n-grams with an overlap of between 1 and 4, based on precision. While ROUGE (Lin, 2004), as an assessment tool, is based on recall, a concept first formally introduced in the context of summaries. Furthermore, a considerable number of neural network-based translation quality assessment methods have been developed in recent times (Guzmán et al., 2017, 2019; Ma et al., 2016; Shao et al., 2024; Gunasekar et al., 2024).

Parallel corpus quality assessment employs sentence-level feature representations, primarily utilised for the execution of corpus filtering tasks. A number of the assessments employ sentences as representatives within the aligned multilingual embedding space, for example, XLM-R (Conneau et al., 2019), MUSE(Conneau et al., 2017), LaBSE(Feng et al., 2020) and LASER(Schwenk and Douze, 2017). There exist alternative works based on pre-trained models or LLM to directly score parallel sentence pairs, e.g., COMET (Rei et al., 2020), NMTscore(Vamvas and Sennrich, 2022), and BERTscore(Zhang et al., 2019).

The aforementioned automated corpus quality assessment methods are at the sentence level, deviating from the token-level conditional probability modelling of the NMT task, which introduces unavoidable noise. Furthermore, the models employed in these approaches are contingent on additional corpora that exhibit considerable variability in their adaptability to different language pairs and scenarios. These corpora are predominantly black-boxed, resulting in uncertainty impacts on the NMT system. Furthermore, extant research has neglected to explore the quantitative relationship between corpus quality and translation quality. The absence of such a quantitative relationship has impeded the execution of in-depth studies of the role of the corpus in NMT systems.

2.2 Neural Machine Translation

The objective of the Neural Machine Translation task is to generate target-side token sequences with the utility of neural networks, given the corresponding source-side token sequences, which are referred to as parallel sentence pairs of the source and target languages. It is possible to divide NMT into two distinct paradigms: the autoregressive paradigm, which generates one token at a step, and the non-autoregressive paradigm, which generates all tokens at one step. The present paper focuses on supervised autoregressive natural language processing (NLP), hereafter referred to as NMT The mathematical modelling of NMT is predicated on the conditional probability of a sequence. Given the source-side token sequence $S = \{s_1, s_2, \cdots, s_J\}$ and the parallel target-side token sequence $T = \{t_1, t_2, \cdots, t_K\}$, the joint probability of the target-side token sequence $P(T \mid S)$ is defined as:

$$P(T \mid S) = \prod_{i=1}^{K} P(t_i \mid S, \hat{T}_{<i}) \quad (1)$$

where $\hat{T}_{<i}$ is the generated target pre-order sequence before t_i, and $P(t_i \mid S, \hat{T}_{<i})$ is the probability of t_i given the source-side token sequence and the generated target-side pre-order sequence. The teacher-forcing strategy, which can avoid the influence of the biased generated pre-order sequence, calculates the probability $P(t_i \mid S, T_{<i})$ with ground truth target pre-order sequence $T_{<i}$.

2.3 Attention-based Networks

The multi-head attention mechanism (Waswani et al., 2017) effectively addressed the long-distance dependency issue and notably enhanced the translation quality of the NMT system, as formalized as follows:

$$\begin{aligned} &\text{MultiHead}(Q, K, V) = \\ &\qquad \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \\ &\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (2)$$

where Q, K, V refer to query, key, and value vectors, W^O is the output projection matrix, W_i^Q, W_i^K, W_i^V are projection matrices of the i-th head, Concat$(\cdot)$ concatenates all heads, Attention$(\cdot)$ computes head$_i$ as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (3)$$

where d_k is the dimension of the key vectors, $\sqrt{d_k}$ is the scaling factor, $softmax$ is the softmax function applied row-wise, and $QK^\top$ calculates the score matrix, which is the core of the attention mechanism.

The flexible score matrix enables the attention mechanism to effectively calculate and extract the association and difference between query and key vectors. In the context of machine learning, a self-attention mechanism that aligns the query and key vectors can facilitate the extraction of internal information from the vector. This process enables the realization of feature extraction. A cross-attention mechanism, in which the query and key vectors are distinct, emphasizes the inter-association and discrepancy, thereby facilitating feature fusion. The multi-head design enables each head to focus on different information independently, thereby enhancing the attention mechanism's feature extraction and expression ability.

3 Methodology

The technical particulars and other prerequisites are delineated before the exposition of the specific methodology. Firstly, this paper adopts the token level as the granularity, i.e., the overall situation of each token in the dictionary over the whole corpus. Token-level granularity is instrumental in ensuring the consistency of NMT mathematical modeling, neural networks, and evaluation metrics. This, in turn, ensures that the study is both rigorous and accurate, while consistent with the holistic nature of quality assessment. Secondly, this paper adopts the teacher-forcing strategy while training and testing the baseline neural network. According to the conditional probability modeling of NMT, the teacher-forcing strategy is strictly consistent with NMT's conditional probability modeling. Moreover, this strategy effectively avoids the accumulation of errors. Thirdly, the Translation Accuracy Prediction (TAP) task is distinct from the Translation Quality Assessment (TQA) and the Translation Quality Estimation (TQE) tasks. TAP utilizes corpus features to predict token-level translation accuracy, thereby modeling the probability distribution between corpus quality and translation quality. The subsequent two tasks are designed to evaluate the quality of the generated translations.

3.1 Corpus Analysis Vector

Corpus Analysis Vector (CAV) is a representation that converts the textual corpus into matrices. The technical core of CAV is the conversion matrix, which converts the text corpus into matrices while following the conditional probability modeling of NMT to retain the necessary information. Under this premise, CAV can simultaneously satisfy the previously mentioned granularity consistency, ensuring the accuracy. The subsequent section will illustrate the formulation of CAV through the transformation process from text corpus to CAV.

According to the conditional probability modeling of NMT, when a neural network is generating a certain target-side token, the probability distribution of the token is determined by its preceding sequence. More specifically, this probability is determined by all the tokens of the source-side sequence and the tokens that precede that token in the target-side sequence. Given the parallel corpus where the source-side dictionary $\mathbf{T}_{\text{src}}$ with N_{src} tokens, the target-side dictionary $\mathbf{T}_{\text{tgt}}$ with N_{tgt} tokens, the CAV is a set of matrices as follows:

$$\begin{aligned} CAV &= \{CAV(t) \mid t \in \mathbf{T}_{\text{tgt}}\} \\ CAV(t) &= \sum_{seq(t) \in \mathbf{SEQ}(t)} \mathrm{F}(t, seq(t)) \end{aligned} \tag{4}$$

where CAV is the matrix set of $CAV(t)$, $\mathbf{SEQ}(t)$ is the set of sequences $seq(t)$ containing token t, and $\mathrm{F}(t, seq(t))$ is the converted matrix of $seq(t)$ in response to t.

Given the parallel sequence pair $\{seq_{\text{src}}(t), seq_{\text{tgt}}(t)\}$, where the target-side sequence is $seq_{\text{tgt}}(t) = \{t_1, t_2, \ldots, t_m\}$ containing token t, the parallel source-side sequence is $seq_{\text{src}}(t) = \{s_1, s_2, \ldots, s_n\}$, the formulation of $\mathrm{F}(t, seq(t)$ is as follows:

$$\begin{aligned} &\mathbf{T}_{\text{c}} = \mathrm{Concat}(\mathbf{T}_{\text{s}}, \mathbf{T}_{\text{t}}) \\ &seq_{\text{c}}(t) = \mathrm{Concat}(seq_{\text{s}}(t), seq_{\text{t}}(t)) \\ &i_{t_i} = \mathrm{Idx}_{\mathbf{T}_{\text{c}}}(t_i),\ \forall t_i \in \mathbf{T}_{\text{c}} \\ &p(t, t_i) = \mathrm{Idx}_{seq_{\text{c}}(t)}(t) - \mathrm{Idx}_{seq_{\text{c}}(t)}(t_i), \\ &\qquad \forall t \in \mathbf{T}_{\text{t}}, t_i \in \mathbf{T}_{\text{c}} \\ &F(t, seq_{\text{c}}(t)) = M \in \mathbb{Z}^{(N_s+N_t)\times L_{\max}} \\ &\forall t_i \in \mathbf{T}_{\text{c}},\ M_{r,c} = \begin{cases} 1, & r = i_{t_i},\ c = |p(t, t_i)|, \\ & p(t, t_i) \geq 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

where $\mathbf{T}_{\text{cat}}$ is the concatenated dictionary of $\mathbf{T}_{\text{src}}$ and $\mathbf{T}_{\text{tgt}}$, $seq_{\text{cat}}(t)$ is the concatenated sequence of $seq_{\text{src}}(t)$ and $seq_{\text{tgt}}(t)$, $\mathrm{Index}_{\mathbf{T}_{\text{cat}}}(t_i)$ is the index of t_i in the concatenated dictionary $\mathbf{T}_{\text{cat}}$, $p(t, t_i)$ is the relative position of t and t_i, $\mathrm{F}(t, seq(t)$ returns the matrix $M \in \mathbb{Z}^{(N_{\text{src}}+N_{\text{tgt}})\times L_{\max}}$ that stores the parallel sequence pair $\{seq_{\text{src}}(t), seq_{\text{tgt}}(t)\}$, where $L_{\max}$ is the maximum length of the concatenated sequences, and $M_{r,c}$ is the element of matrix M in the r-th row and c-th column.

In this way, M stores the identity and relative position information of all the preceding sequence tokens of the parallel sequence pair $\{seq_{\text{src}}(t), seq_{\text{tgt}}(t)\}$ in response to token t, i.e., all the source sequence tokens and the target sequence tokens before t. Thus, the $CAV(t)$ is obtained by accumulating all the matrices M corresponding to the parallel sequence pairs $\{seq_{\text{src}}(t), seq_{\text{tgt}}(t)\}$ in response to t At this juncture, the CAV has successfully stored the requisite information, which includes pertinent data from which the corpus quality can be induced.

3.2 Attention-based Network For Translation Accuracy Prediction

As previously stated, CAV is a collection of corpus features for each token of the target side, where there are intricate relationships among the corpus features of each token. Consequently, the proposed neural network is designed to mine internal relationships, which is precisely the strength of the attention mechanism. Furthermore, given the common dictionary scale of NMT tasks, direct utilization of the attention mechanism necessitates substantial resource consumption, necessitating dimensionality reduction on CAV as a prerequisite.

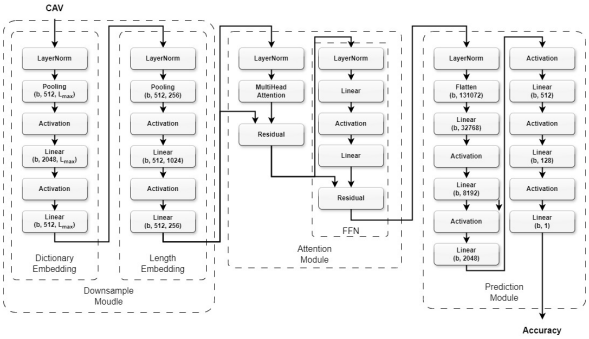

Figure 1: Architecture of ANTAP

As shown in Fig. 1, the Attention-based Network For Translation Accuracy Prediction (ANTAP) consists of three primary components, namely, the

downscale module, the attention module, and the prediction module.

The original input CAV_{in} has a dimension of batch $\times (N_{\text{src}} + N_{\text{tgt}}) \times L_{\text{max}}$, where the first dimension refers to the batch size, the second refers to the concatenated dictionary scale, the third refers to the maximum concatenated sequence length. ANTAP employs the adaptive mean pooling technique, which involves the downsampling of the final two dimensions of the CAV. The subsequent embedding modules perform a straightforward information integration of the final two dimensions of the CAV, each responsible for embedding the downsampled dimensions. The following is the formulation of the aforementioned downscale module:

$$\begin{aligned} CAV_{\text{ds}} &= \text{GELU}\Big(\text{AdaptiveAvgPool2D}\big(\\ &\qquad \text{LayerNorm}(CAV_{\text{in}})\big)\Big) \\ CAV_{\text{emb}} &= \text{EmbLinear}^{\text{2nd}}\Big(\\ &\qquad \text{GELU}\big(\text{EmbLinear}^{\text{3rd}}(CAV_{\text{ds}})^{\text{T}}\big)\Big) \end{aligned} \tag{6}$$

where CAV_{ds} is the downsampled vector, AdaptiveAvgPool2D$(\cdot)$ controls the downsampling factor, EmbLinear$^{\text{2nd}}(\cdot)$ and EmbLinear$^{\text{3rd}}(\cdot)$ transformation of the second and third dimensions of CAV_{ds}, respectively.

The embedded vector is then fed into the Multihead Attention Module, which is designed as referred to (Waswani et al., 2017) to extract the intricate relationships among the tokens, as follows:

$$\begin{aligned} CAV'_{\text{attn}} &= \text{Multihead}\Big(\text{LayerNorm}(CAV_{\text{in}})\Big) \\ &\quad + CAV_{\text{in}} \\ CAV_{\text{attn}} &= \text{FFN}\Big(CAV'_{\text{attn}}\Big) \end{aligned} \tag{7}$$

where CAV'_{attn} is the intermediate variable of CAV_{attn}, and FFN$(\cdot)$ consists of linear layers and the activation function as referred to (Waswani et al., 2017).

The multi-attention mechanism's long-range performance facilitates the integration of features representing complex relationships among tokens that interact with each other into CAV_{attn}.

Considering the token-level translation accuracy as a scalar, the prediction module of the neural network needs to integrate the features extracted by the attention module. In particular, the prediction module of the network consists of alternating linear layers and activation functions as follows:

$$CAV_{\text{flat}} = \text{PredModule}(\text{Flatten}(CAV_{\text{attn}})) \tag{8}$$

where Flatten$(\cdot)$ flattens the last two dimensions of CAV_{attn}, PredModule$(\cdot)$ downsamples the flattened vector step by step until reduced to 1. Consequently, ANTAP attains token-level translation accuracy by leveraging token-level CAV as the input.

3.3 Translation Accuracy Prediction Task

The subsequent section delineates the Translation Accuracy Prediction (TAP) task, which is designed to model the quantitative relationship between corpus quality and translation quality. The TAP task, therefore, aims to predict token-level translation accuracy, operating under the assumption that the given baseline network has been adequately trained. Under this consideration, the discrepancy in accuracy between tokens is mainly attributable to the probability distribution of the parallel corpus. Given the corpus features $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N\}$, the mathematical modeling of the TAP task is as follows:

$$P(\text{Acc} \mid \mathcal{C}) = \prod_{i=1}^{N} P\left(\text{acc}_i \mid \mathbf{c}_i, \mathcal{F}(\mathcal{C})\right) \tag{9}$$

where $\mathbf{c}_i \in \mathbb{R}^d$ is the corpus feature of the i-th token, $\mathcal{C}$ determines the token-level accuracy $\text{Acc} = \{\mathbf{acc}_1, \mathbf{acc}_2, \ldots, \mathbf{acc}_N\}$ with the probability distribution, and $\mathcal{F}(\mathcal{C})$ is the quantitative relationship between corpus features and translation accuracy.

4 Experiment

4.1 Dataset And Settings

The datasets selected for this study are drawn from the publicly accessible datasets, namely WMT14 English-German (En-De), WMT17 English-Chinese (En-Zh), WMT21 German-Upper Sorbian (De-Hsb), and WMT21 Russian-Chuvash (Ru-Chv). These corpora encompass a wide range of scenarios, from those with limited resources to those with abundant resources, and span multiple language pairs. The pre-processing stage involves a series of essential steps, including the implementation of blank line filtering, special symbol filtering, length filtering, lower-case conversion, tokenization, and sub-word processing, as outlined in (Ott et al., 2019). Depending on the experimental

Table 1: Baseline BLEU Scores

	WMT14	WMT17	WMT21	
	En-De	En-Zh	De-Hsb	Ru-Chv
BLEU	26.77	21.01	47.72	17.70

platform, the pre-processing filters out sub-word sequences longer than 128, and sets the source-side and target-side maximum sub-word dictionary scale to 40k.

The baseline translation network is trained to model the probability distribution of the training set. Thus, the paper explores the relationship between corpus quality and translation quality on the training set. It is imperative that a subset partitioning of the target-side dictionary is performed, entailing the random division of the target-side dictionary into training, development, and testing subsets in a ratio of 8:1:1. The ANTAP was trained with MSE as the objective, with the AdamW optimizer. The ANTAP model was evaluated using the following metrics: MSE, MAE, RMSE, and R^2. The experimental platform employs the Ubuntu 22.04 operating system, accompanied by 512GB of RAM, utilising an A6000 GPU with 48GB of memory.

4.2 Baseline Translation Network

The baseline model selected for this experiment adopts the basic Transformer model (Waswani et al., 2017), where the parameter configuration is consistent with that employed in the original paper. As previously mentioned, the baseline translation network was trained and tested using the teacher-forcing strategy and evaluated with SacreBLEU, as illustrated in the subsequent table. As demonstrated in Table 1, the baseline NMT system has attained the performance levels of other baseline NMT systems documented in the literature on the four corpora, which indicates that the models have been adequately trained.

4.3 Translation Accuracy Prediction

The experimental results of ANTAP on the TAP task are hereby exhibited in two states. State I: Optimal checkpoint before the early-stop is triggered during the training process, where the ANTAP can be regarded as attaining its optimal performance on the test set. State II: Last checkpoint when the early-stop is triggered during the training process,

Table 2: Training subsets scores of State I

	WMT14	WMT17	WMT21	
	En-De	En-Zh	De-Hsb	Ru-Chv
MSE	0.0002	0.0016	0.0120	0.0001
MAE	0.0102	0.0147	0.0758	0.0060
RMSE	0.0149	0.0400	0.1098	0.0085
R^2	0.9967	0.9823	0.5866	0.9991

Table 3: Test subsets scores of State II

	WMT14	WMT17	WMT21	
	En-De	En-Zh	De-Hsb	Ru-Chv
MSE	0.0253	0.0108	0.0192	0.0397
MAE	0.1155	0.0356	0.0922	0.1511
RMSE	0.1589	0.1040	0.1384	0.1992
R^2	0.6246	0.8824	0.3639	0.5093

where the ANTAP can be regarded as attaining its optimal performance on the training set. Note that to investigate the convergence and generalizability of CAV-based ANTAP, ANTAP was deliberately designed to eliminate the dropout mechanism. The ensuing tables present the mean squared error (MSE), the mean absolute error (MAE), the root mean squared error (RMSE), and the R-squared (R^2) of ANTAP on the training and test subsets of the four corpora.

As illustrated in Table 2, the performance of CAV-based ANTAP on the training set of State I is demonstrated. The table demonstrates the efficacy of CAV-based ANTAP in terms of accuracy and generalisation when applied to the WMT14 En-De, WMT17 En-Zh, and WMT21 Ru-Chv corpora. Additionally, ANTAP is capable of explaining more than 98% of the variation in the dependent variable. Despite achieving marginally lower scores, ANTAP still achieves considerable accuracy and generalisation on the WMT21 De-Hsb corpus, where ANTAP is able to explain more than half of the variance in the dependent variable.

As illustrated in Table 3, the performance of CAV-based ANTAP on the test set of State I is demonstrated. In comparison with the training set of State I, ANTAP demonstrates lower levels of accuracy and generalisation on the test set for all four corpora. The performance of the test set on the WMT17 En-Zh corpus is the closest to the training set. Nevertheless, ANTAP continues to

Table 4: Training subsets scores of State II

	WMT14	WMT17	WMT21	
	En-De	En-Zh	De-Hsb	Ru-Chv
MSE	0.0000	0.0001	0.0000	0.0000
MAE	0.0014	0.0020	0.0033	0.0005
RMSE	0.0022	0.0083	0.0069	0.0008
R^2	0.9999	0.9992	0.9984	1.0000

Table 5: Test subsets scores of State II

	WMT14	WMT17	WMT21	
	En-De	En-Zh	De-Hsb	Ru-Chv
MSE	0.0254	0.0112	0.0232	0.0414
MAE	0.1154	0.0341	0.0949	0.1526
RMSE	0.1593	0.1061	0.1523	0.2034
R^2	0.6229	0.8777	0.2298	0.4884

demonstrate commendable accuracy and capacity for generalisation.

As illustrated in Table 4, the performance of CAV-based ANTAP on the training set of State II is demonstrated. The table demonstrates the efficacy of CAV-based ANTAP in terms of accuracy and generalisation when applied to all four corpora. The analysis of these corpora reveals that ANTAP is capable of explaining more than 99% of the variation in the dependent variable.

As illustrated in Table 5, the performance of CAV-based ANTAP on the test set of State II is demonstrated. In comparison with the training set of State II, ANTAP demonstrates lower levels of accuracy and generalisation on the test set than on the training set for all four corpora. The performance of the test set on the WMT17 En-Zh corpus is the closest to that of the training set. In comparison with the test set of State I, ANTAP demonstrates slightly lower levels of accuracy and generalisation on all four corpora. Nevertheless, ANTAP continues to demonstrate commendable accuracy and capacity for generalisation.

In summary, the CAV-based ANTAP demonstrates impressive accuracy and fitting capabilities. Furthermore, CAV-based ANTAP exhibits favourable generalisation capabilities. Up to this point in the study, the paper has effectively established a quantitative relationship between the corpus quality and the translation quality, which has been achieved by constructing probability distributions for CAV and translation accuracy.

4.4 Case Study

In terms of the translation quality, tokens characterised by low translation accuracy are of greater concern than tokens characterised by high translation accuracy. The present section thus aims to conduct a case study with CAV in order to analyse the causes of the poor translation accuracy. Following the conditional probability model of NMT, it is theorised that the underperformance of specific tokens is attributable to the presence of analogous tokens within the antecedent sequence of the former, i.e., the CAV exhibits similarity. In order to validate the hypothesis, the low-precision token in the target-side dictionary is first targeted, and then the token with the highest cosine similarity, referred to as the suspect tokens, to the CAV of the targeted token is located. The following table exhibits the target and suspect tokens for the four corpora, along with their cosine similarity and accuracy.

As illustrated in Table 6, the four corpora are divided into three columns, from left to right, denoting the token, the cosine similarity, and the token-level accuracy of the low-accuracy tokens and suspect tokens. The suspect tokens exhibit a high degree of cosine similarity to the CAV of the target token. However, no discernible correlation is observed between the translation accuracy of the target and suspect tokens. It is hypothesised that, under the conditional probability model of NMT, translating tokens with a low degree of CAV difference would be considerably more difficult. The translation accuracy of these tokens is demonstrated to exhibit stochasticity and instability. Supposing the tokens have the same CAV, i.e., in the case of a one-to-many linguistic phenomenon, instead of generating these two tokens with equal probability, the NMT system generates them randomly while maintaining the probability sum of these two tokens constant.

To test the aforementioned hypotheses, ablation experiments of CAV modifications were performed in this paper. Specifically, the CAV is modified by appending an artificial pseudo-token to the source-side sequence tail of the targeted token to distinguish it from the suspect token. The subsequent table illustrates the cosine similarity and accuracy of target and suspect tokens in the NMT system that has been trained using the modified corpus.

As illustrated in Table 7, the CAV similarities

Table 6: Original results of target and suspect token (Token/Accuracy)

	WMT14		WMT17		WMT21			
	En-De		En-Zh		De-Hsb		Ru-Chv	
Target Token	gesamteuropäi@@	0.0000	vo	0.0116	prašeše	0.0000	л е к н и с е н е	0.0000
Similar Token	anstehenden	0.2269	翔@@	0.1645	jeju	0.4348	п у @@	0.3173
Cosine Similarity	0.9857		0.9314		0.8420		0.8732	

Table 7: Ablation results of target and suspect token (Token/Accuracy)

	WMT14		WMT17		WMT21			
	En-De		En-Zh		De-Hsb		Ru-Chv	
Target Token	gesamteuropäi@@	0.4016	vo	0.7011	prašeše	0.8125	л е к н и с е н е	0.6909
Similar Token	anstehenden	0.2409	翔@@	0.1865	jeju	0.5652	п у @@	0.3035
Cosine Similarity	0.7953		0.7591		0.4498		0.6907	

of the target and suspect tokens are decreased. In addition, the accuracy of the target tokens improves, whereas there is no clear pattern in the change in the accuracy of the suspect tokens. The increase in target token accuracy is attributed to the modified CAV of the target token has been differentiated from the suspect token, while the uncertain change in the accuracy of suspicious tokens is due to the effect of other tokens in the corpus.

The ablation experiments confirmed our hypothesis about the relationship between CAV similarity and low translation accuracy.

5 Conclusions

In this paper, we propose a data-driven performance-oriented corpus quality assessment tool, a translation accuracy prediction task, and a corresponding network based on the attention mechanism. Based on CAV and TAP tasks, ANTAP successfully modelled the probability distribution of CAV and translation accuracy, and established the quantitative relationship between corpus quality and translation quality. Besides, this paper demonstrates the impressive capabilities of CAV-based ANTAP in corpus quality assessment and analysis, and also shows the great potential of CAV in downstream applications, i.e., corpus filtering.

6 Discussion

The corpus quality assessment method proposed in this paper maintains a high level of consistency by strictly following the conditional probability modelling of NMT from the token-level perspective, which is different from the existing work at the sentence level. Consequently, the methodology outlined in this paper has exhibited a high degree of efficacy in experimental settings. However, it should be noted that the methodology of the paper is not without limitations, which represent a direction for future work. Firstly, it is evident that CAV, CAV-based ANTAP, and TAP tasks all rely on a trained baseline model, which imposes limitations on the application scenarios that can be utilised. Consequently, the utilisation of unsupervised methodologies founded upon CAV emerges as a particularly auspicious research domain. Second, the macroscopic properties of CAV can introduce additional noise, such as an illusory corpus that is not present in the corpus, but which conforms to CAV. While the impact of these potential errors is deemed to be negligible in terms of the experimental results in this paper, they are nevertheless worthy of note. In conclusion, it can be argued that the corpus quality assessment and corpus analysis paradigm outlined in this study have considerable potential in the field of NMT.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ran-

zato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

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

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Conti Kauffmann, Gustavo Henrique de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Behl, Xin Wang, Sebastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2024. Textbooks are all you need.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2017. Machine translation evaluation with neural networks. *Computer Speech & Language*, 45:180–200.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2019. Pairwise neural machine translation evaluation. *arXiv preprint arXiv:1912.03135*.

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

Qingsong Ma, Fandong Meng, Daqi Zheng, Mingxuan Wang, Yvette Graham, Wenbin Jiang, and Qun Liu. 2016. Maxsd: A neural machine translation evaluation metric optimized by maximizing similarity distance. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*, pages 153–161. Springer.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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*, pages 311–318.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jannis Vamvas and Rico Sennrich. 2022. Nmtscore: A multilingual analysis of translation-based text similarity measures. *arXiv preprint arXiv:2204.13692*.

A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *NIPS*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.