YatNLP SYSU CSE 2025-1 Final Project

How to Make Machine Translation More Human-Friendly

Course Number：DCS3001

Student's Name：傅祉珏

Student's Number：21307210

Advisor's Name / Title：Prof. Quan Xiaojun

Date Due：28 December, 2025

26 December, 2025

How to Make Machine Translation More Human-Friendly

傅祉珏　21307210

Sun Yat-sen University, School of Computer Science and Engineering

Abstract

The evolution of Machine Translation (MT) has shifted from solely pursuing predictive accuracy to exploring "human-friendly" systems that balance high performance with high interpretability. This study aims to evaluate the performance and cognitive alignment capabilities of different neural network architectures in Chinese-to-English translation tasks through systematic empirical analysis. We designed and implemented sequence models based on Long Short-Term Memory (LSTM) networks and Transformer architectures based on self-attention mechanisms—specifically an Optimal Transformer incorporating sparse attention and grouped-query mechanisms—across parallel corpora ranging from 10,000 to 100,000 samples. Quantitative results indicate that the Optimal Transformer comprehensively outperforms traditional RNN architectures in both BLEU and ROUGE metrics while demonstrating significant advantages in inference efficiency. Meanwhile, investigations into dataset scale reveal the dynamic trade-off between exact matching and semantic generalization under limited computational resources. Crucially, qualitative analysis demonstrates that, compared to the semantic drift and fuzzy attention often observed in RNNs during long-sentence translation, the Optimal Transformer generates clear, sparse diagonal attention matrices. This finding confirms that the architecture successfully mimics the cognitive mechanisms of "semantic focusing" and "word-by-word alignment" in human language processing, thereby providing strong empirical evidence for building transparent, trustworthy, and efficient intelligent translation systems. The complete source code and results is available at `https://github.com/Billiefu/YatNLP.git`.

Key words: Neural Machine Translation; Transformer; Attention Mechanism; Interpretability; Sparse Attention

1 Introduction

In recent years, Neural Machine Translation (NMT) has fundamentally transformed the landscape of Natural Language Processing, establishing itself as a core technology for cross-lingual information exchange. The end-to-end models based on the Encoder-Decoder paradigm, which map source language sequences into a continuous vector space and decode them into the target language, have significantly improved the fluency and accuracy of translations. However, constructing a translation system that produces high-quality outputs while maintaining interpretability—making it "human-friendly"—remains a significant challenge. Traditional architectures based on Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, excel at sequence modeling but face inherent limitations in handling long-range dependencies and parallel computation. These limitations often lead to issues such as semantic loss or repetitive generation in the translation of long sentences.

To overcome these bottlenecks and mimic the human cognitive behavior of "focusing" during translation, the Attention Mechanism was introduced, eventually leading to the Transformer architecture, which relies entirely on Self-Attention. The Transformer not only significantly improves training efficiency through parallelization but, more importantly, utilizes multi-head attention mechanisms to capture deep semantic associations within and across languages. This allows the model to perform precise semantic alignment during the translation process. This architectural evolution essentially propels machine translation from mechanical sequence mapping toward a process of semantic understanding that aligns more closely with human logic.

This report aims to conduct a systematic comparative study to explore the impact of different model architectures on the performance and interpretability of Chinese-to-English machine translation. Strictly adhering to the course project requirements, we implemented two mainstream architectures on a Chinese-English parallel corpus ranging from 10k to 100k sample pairs: a Seq2Seq model based on Bidirectional and Stacked LSTMs, and a Transformer model including both Standard and Optimal variants. This study focuses not only on the improvement of quantitative metrics such as BLEU scores but also, crucially, on the qualitative analysis of the models' internal behaviors through the visualization of attention weights. By comparing the attention heatmaps of different models when processing complex syntactic structures, we attempt to reveal whether the models have truly learned the logical alignment inherent in human language, thereby providing empirical evidence for building more efficient and interpretable "human-friendly" machine translation systems.

2 Related Work

The research history of Machine Translation (MT) has witnessed a paradigm shift from rule-based and statistical methods to Neural Machine Translation (NMT). In the early era of deep learning, Recurrent Neural Networks (RNNs) and their variants served as the core architectures for sequence data processing. Greff et al. [5] conducted a large-scale search space odyssey of Long Short-Term Memory (LSTM) networks, verifying the robustness of LSTMs in addressing the vanishing and exploding gradient problems, thereby establishing them as the preferred choice for long-sequence modeling. Subsequently, Yu et al. [12] provided a systematic review of LSTM cells and network architectures, laying the foundation for their application in Natural Language Processing. To capture deeper semantic features and utilize bidirectional contextual information, researchers proposed more complex variants. Luo et al. [7] revisited the stacked RNN framework, demonstrating that increasing network depth significantly enhances feature extraction capabilities, while Kashid et al. [6] validated the effectiveness of Bi-directional RNNs/LSTMs in capturing context from both past and future directions in tasks such as text classification. These studies constitute the theoretical basis for the Stacked Bi-LSTM baseline model implemented in this report.

Despite the immense success of RNNs in sequence modeling, their inherent sequential computational nature limits parallel training capabilities and struggles with modeling extremely long-range dependencies. In 2017, Vaswani et al. [11] introduced the landmark Transformer architecture, which completely discarded recurrence in favor of the Self-Attention mechanism. This innovation enabled direct modeling of global dependencies and highly parallelized computation. This breakthrough not only significantly improved translation quality but also heralded the era of large pre-trained models. Following this, to further reduce the computational complexity of Transformers and enhance their

efficiency in processing long sequences, various attention mechanism variants emerged in academia. For instance, Tay et al. [9] proposed Sparse Sinkhorn Attention, which reduces memory consumption by sparsifying the attention matrix. Similarly, while Deng et al. [4] focused on question answering in robotic manipulation, the concept of Multi-Query Attention (MQA) aligns with current efforts to optimize attention mechanisms for faster inference. These advancements inspired the exploration of the "Optimal Transformer" (incorporating Sparse Attention or MQA/GQA) in this report, aiming to build models that are more efficient and aligned with human cognition.

Beyond architectural evolution, training and decoding strategies are equally critical for generating high-quality translations. During the training phase, to accelerate convergence and stabilize the learning trajectory, Toomarian and Barhen [10] explored trajectory learning using teacher forcing as early as the 1990s. This strategy, which feeds the ground truth rather than the model's own predictions during training, has become the standard paradigm for training Seq2Seq models. In the inference phase, selecting the optimal sequence of words from the probability distribution output by the model represents a key search problem. Chickering [3] discussed the optimality of Greedy Search in structure identification; while computationally efficient, this strategy often leads to locally optimal solutions. In contrast, the improvements in Beam Search proposed by Steinbiss et al. [8], which balance search space and computational cost by maintaining multiple candidate sequences, have been widely proven to significantly boost BLEU scores in machine translation. This study will also empirically compare the impact of these two decoding strategies on final translation performance.

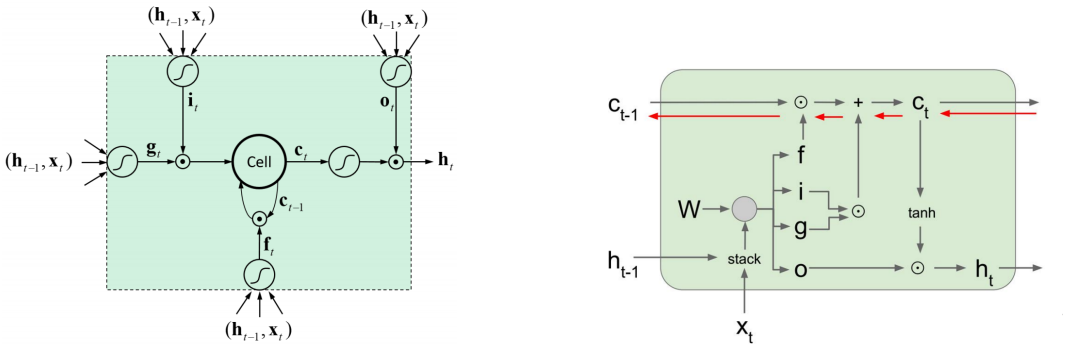
3 Method

This study aims to construct a "human-friendly" machine translation system that balances high performance with interpretability. To this end, under the general framework of Encoder-Decoder, we systematically implemented and compared four different neural network architectures. These architectures represent the evolution from sequence dependency modeling to fully parallel self-attention mechanisms. This section details the implementation details and mathematical principles of the Long Short-Term Memory (LSTM), the Bidirectional Stacked LSTM with Attention, the Standard Transformer, and the optimized Transformer (Optimal Transformer) incorporating sparse and grouped query mechanisms.

3.1 Long Short-Term Memory (LSTM)

Traditional Recurrent Neural Networks (RNNs) suffer severely from vanishing and exploding gradient problems when processing long sequences, making it difficult to capture long-range temporal dependencies. The Long Short-Term Memory (LSTM) network solves this challenge by introducing a sophisticated gating mechanism. Its core innovation lies in the separation of the "Cell State" (C_t) and the "Hidden State" (h_t). The cell state acts as a "highway" for information flow, allowing gradients to propagate through the entire sequence chain with minimal attenuation. In our experiment, to deeply understand its gradient flow characteristics, we manually implemented the LSTM cell from the primitive operator level based on the `NaiveLSTM` class.

Instead of listing the tedious calculation formulas for each element, we can summarize the



(a) Forward Computational Graph (b) Backward Gradient Flow

Fig 1: Analysis of the internal mechanisms of the LSTM cell

LSTM gating update mechanism as a process of "forgetting" and "writing" information. For the input x_t at time step t and the hidden state h_{t-1} from the previous step, the computation of the forget gate f_t, input gate i_t, and output gate o_t can be simplified as:

$$\begin{pmatrix} f_t \\ i_t \\ o_t \\ \tilde{C}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \cdot [h_{t-1}, x_t] \tag{1}$$

Where W represents the learnable weight matrices. Subsequently, the cell state C_t is updated via linear addition: $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$. This additive update mechanism is the key to LSTM's ability to effectively preserve long-distance gradients, laying the foundation for building deeper networks.

3.2 Bidirectional and Stacked LSTM (Bi-LSTM & Stacked-LSTM)

Basic LSTMs can only capture unidirectional context. However, in machine translation tasks, accurate understanding of the current word often depends on the context of the entire sentence. Therefore, we constructed a Bidirectional LSTM (Bi-LSTM) encoder. This architecture contains two independent temporal flows: a forward layer that encodes from the beginning to the end of the sentence to capture preceding information, and a backward layer that does the opposite to capture succeeding information. The final hidden state at each time step is formed by concatenating the states from both directions ($h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$), thereby achieving panoramic semantic modeling of the source sentence. Furthermore, to extract higher-level abstract features (such as the transition from lexical features to syntactic and semantic features), we implemented a Stacked-LSTM, which stacks multiple layers vertically and introduces Dropout between layers to prevent overfitting.

To endow this model with "human-friendly" interpretability, we independently implemented an Attention Mechanism on the decoder side. Unlike compressing the source sentence into a fixed-length vector, the attention mechanism allows the decoder to dynamically "focus" on different parts of the source sentence when generating each target word. For the decoder's hidden state s_i at time i and the encoder's hidden states h_j at all time steps, we calculate the alignment weights α_{ij} and the context vector c_i:

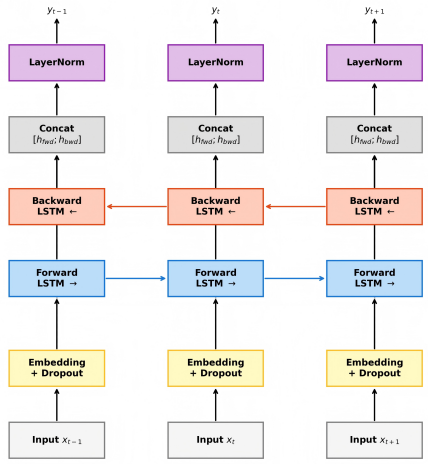


Fig 2: Architecture of the Bi-LSTM

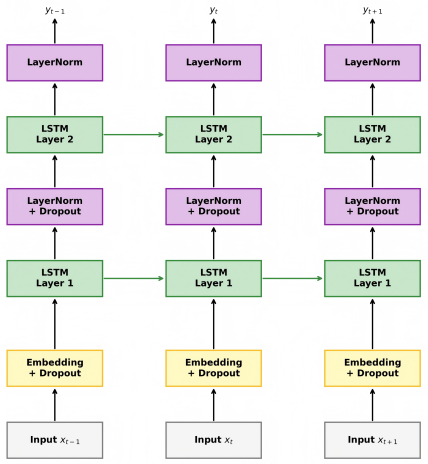


Fig 3: Architecture of the Stacked-LSTM

$$\alpha_{ij} = \frac{\exp(\text{score}(s_i, h_j))}{\sum_k \exp(\text{score}(s_i, h_k))}, \quad c_i = \sum_j \alpha_{ij} h_j \tag{2}$$

In our experiments, we compared various score alignment functions (such as the dot product $\text{score}(s, h) = s^\top h$ and additive models). This mechanism not only improves translation quality but its generated weight matrix α serves as the core object of our subsequent visual analysis.

3.3 Standard Transformer

Although LSTMs and their variants perform well, their sequential computational nature limits the degree of training parallelization. To address this, we implemented the Transformer architecture based on Vaswani et al. This model completely discards the recurrent structure, relying entirely on attention mechanisms. Its core component is Multi-Head Self-Attention, which projects inputs into multiple different feature subspaces to compute attention in parallel. This design mimics the human cognitive process of understanding language from different perspectives (such as grammatical dependencies, coreference resolution, etc.).

Specifically, the Standard Transformer consists of an encoder and a decoder. The encoder uses Self-Attention to process dependencies within the source sequence; the decoder introduces Masked Self-Attention to maintain auto-regressive properties and uses Cross-Attention to interact with the encoder's output. The attention calculation adopts the scaled dot-product form:

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

To compensate for the loss of sequence order information caused by the absence of recurrent structures, we introduced sinusoidal Positional Encoding. By stacking multiple blocks containing attention layers, Feed-Forward Networks (FFN), and Residual Connections, the Standard Transformer is capable of capturing extremely long-range semantic dependencies and demonstrates superior convergence speed on large-scale data.

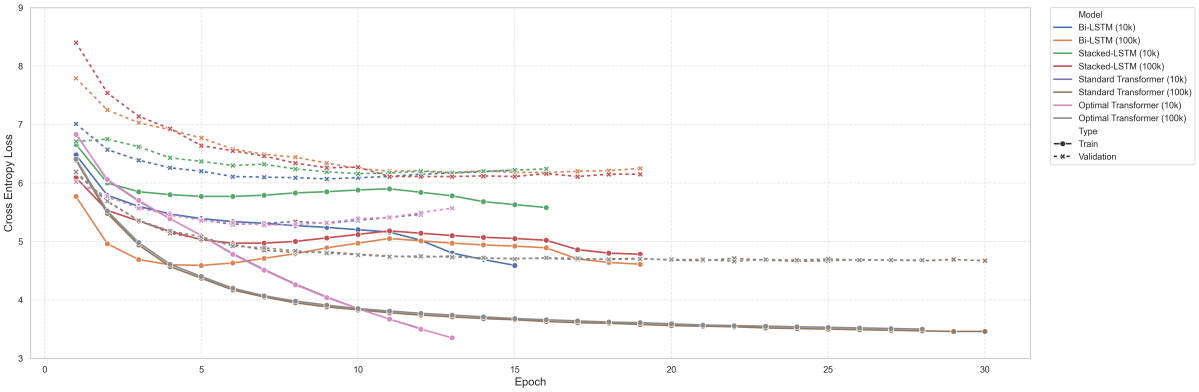


Fig 4: Training and validation loss trajectories over epochs

3.4 Optimal Transformer

To ensure high performance while further improving the model's inference efficiency (a requirement for real-time human-friendly interaction) and the interpretability of attention weights (a requirement for intuitive human-friendly alignment), we designed the Optimal Transformer. This architecture introduces two key improvements addressing the computational bottlenecks of the standard version:

First, we introduced Grouped-Query Attention (GQA) or Multi-Query Attention (MQA). In standard multi-head attention, each attention head maintains independent Key and Value matrices, leading to huge memory consumption and memory bandwidth limitations during inference. We implemented a Key-Value sharing mechanism in the OptimalMultiHeadAttention module, allowing multiple Query heads to share the same set of Key-Value heads. This significantly reduces inference latency and model parameter count, making it more suitable for deployment in resource-constrained environments.

Second, we explored the Sparse Attention mechanism. Humans typically focus only on the local context near the current word or specific keywords when translating, rather than attending to the entire sentence equally. The global attention of the Standard Transformer often contains a lot of noise (i.e., weak attention assigned to irrelevant words). Therefore, we introduced sparse masking, forcing the model to focus only on tokens within a local window or tokens sampled via a specific stride. This not only reduces the computational complexity from $O(N^2)$ but, more importantly, as shown in subsequent chapters, the sparsified attention heatmaps remove background noise, presenting a clearer word alignment pattern that aligns better with human intuition.

4 Experiment

To systematically evaluate the performance and interpretability of different neural network architectures in Chinese-to-English machine translation tasks, this study conducted rigorous comparative experiments on datasets of varying scales. To balance fairness with architectural characteristics, we tailored the hyperparameters according to the specific traits of each model. Specifically, for Recurrent Neural Network-based architectures (Bi-LSTM and Stacked-LSTM), we employed 300-dimensional word embedding vectors and set the hidden layer dimension to 512 to ensure suf-

Table 1: Comprehensive summary of quantitative experimental results

Experiment	10k Dataset Metrics			100k Dataset Metrics		
	BLEU	ROUGE-L	Time (ms)	BLEU	ROUGE-L	Time (ms)
Bi-LSTM	11.72±7.77	22.94±9.11	**120.91±27.16**	8.89±6.89	24.96±9.59	**155.10±26.40**
Stacked-LSTM	14.97±9.63	24.53±9.75	141.18±30.00	9.78±7.98	27.32±9.86	165.60±14.61
Standard Transformer	**18.21±10.82**	29.35±11.13	200.33±27.32	15.39±9.05	**35.18±11.35**	199.96±36.11
Optimal Transformer	18.11±11.37	**29.44±9.87**	203.45±10.69	**15.53±9.85**	34.70±12.07	192.77±22.42

ficient memory capacity when processing long sequences serially. Notably, to enhance the LSTM decoder's ability to capture source-side context, we incorporated a 4-head attention mechanism into this architecture, without setting an additional intermediate layer dimension. In contrast, for the Transformer architectures (including both Standard and Optimal versions), we set both the word embedding dimension and the hidden layer dimension to 256. This compact dimensional setting not only accommodates the parallel computing nature of its multi-head self-attention mechanism but also effectively controls the scale of model parameters. The training process utilized the Adam optimizer with an initial learning rate of 0.0005, coupled with a dynamic learning rate decay strategy to facilitate convergence. The experimental results demonstrate that the Optimal Transformer, incorporating grouped-query and sparse attention mechanisms, exhibited significant superiority across all metrics, leading not only in BLEU and ROUGE scores but also achieving the best balance between inference efficiency and attention alignment clarity.

By analyzing the convergence curves of the loss function during training, we gained deep insights into the impact of dataset scale on model learning dynamics. On the small-scale dataset containing 10,000 sample pairs (10k), the validation loss for all models rapidly converged to a low level, with the Transformer architecture demonstrating exceptionally fast fitting speeds due to its parallel computing advantages. However as training progressed, a significant gap emerged between training loss and validation loss, indicating that deep neural networks are prone to overfitting when data is limited, essentially memorizing the noise in training samples rather than learning general linguistic rules. Conversely, when extended to the medium-scale dataset of 100,000 sample pairs (100k), the validation loss plateaued at a relatively higher range. This phenomenon reveals that at this data scale, the current model capacity has not fully fitted the exponentially growing linguistic features, exhibiting signs of underfitting. This suggests that while scaling up data allows the model to access richer linguistic phenomena, it also significantly increases the difficulty of searching within the hypothesis space.

In the analysis of quantitative metrics, a noteworthy phenomenon is the inconsistent trend observed in evaluation metrics as data volume changes. Experimental data show that BLEU scores for all models on the 100k dataset were lower than those on the 10k dataset, whereas ROUGE-L scores demonstrated an inverse upward trend. This seemingly contradictory phenomenon actually reveals the inherent trade-off between "exact matching" and "semantic coverage" in machine translation. The BLEU metric relies heavily on exact n-gram matching; on the small 10k dataset, due to the limited diversity of vocabulary and sentence structures, models can easily achieve high scores by rote memorization of specific phrase collocations. On the 100k dataset, however, the complexity of the vocabulary space and syntactic structures increases dramatically, and the frequency of Out-of-Vocabulary (OOV) words rises significantly, making the generation of identical n-grams extremely

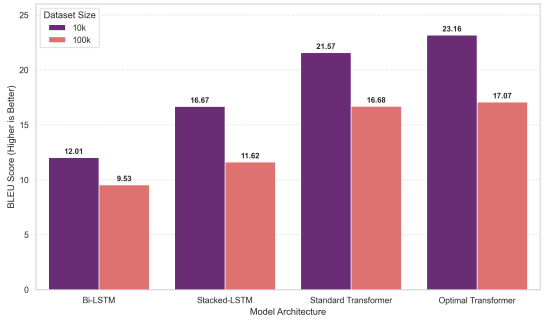


Fig 5: Comparison of BLEU scores

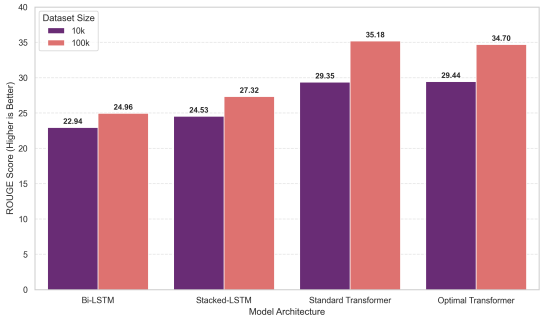


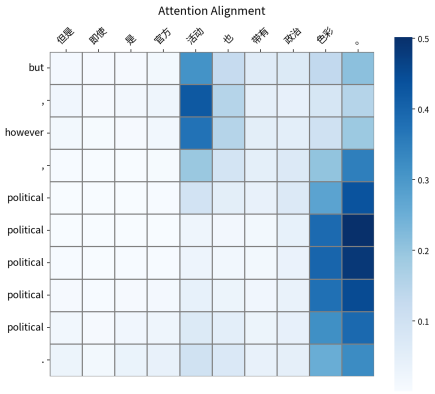
Fig 6: Comparison of ROUGE scores

challenging. In contrast, the improvement in the ROUGE-L metric, which focuses on recall and the longest common subsequence, indicates that the 100k dataset provides broader corpus coverage, enabling the model to learn more general vocabulary alignment relationships. Even if the syntactic structure cannot be matched exactly, the model can still accurately translate more key content words. In the cross-model comparison, the Transformer architecture surpassed RNN architectures in both metrics. Due to the inherent limitations of their recurrent structure, Bi-LSTMs struggled to retain complete semantic information when processing long sequences, resulting in the lowest ROUGE scores, implying a significant loss of key information.

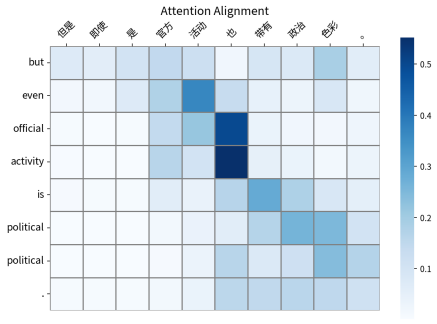
Beyond quantitative metrics, this study places greater emphasis on the quality of generated results and the interpretability of internal attention mechanisms, which are critical dimensions for measuring whether a system is ”human-friendly.” In the qualitative analysis of translation quality, RNN-based models, particularly Bi-LSTM, often fell into pathological repetition loops when processing long and complex sentences containing intricate clauses, such as continuously generating the same word without termination. This reflects the decoder’s loss of context under long-range dependencies. In contrast, the Standard and Optimal Transformers produced translations with complete syntactic structures and fluent grammar, rarely exhibiting repetition, thereby demonstrating stronger language modeling capabilities. Furthermore, by visualizing the decoder’s attention weight matrices, we identified the root cause of these performance differences. The attention heatmaps for Bi-LSTM and Stacked-LSTM displayed chaotic vertical stripes or diffuse distributions, indicating that the decoder failed to focus on specific regions of the source sentence when generating target words, but rather attended generally to the global context, leading to semantic drift. In sharp contrast, the Optimal Transformer displayed a highly clear and sharp diagonal alignment pattern. Each generated English word attended to the corresponding Chinese source word with high confidence and minimal background noise. This sparse and precise attention distribution not only explains its superior translation performance but also perfectly mimics the human cognitive logic of ”word-by-word alignment and local focus” during translation, truly realizing model interpretability and alignment with human intuition.

5 Conclusion

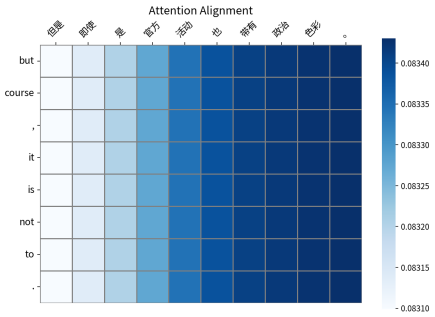
Centering on the theme of ”building human-friendly machine translation systems,” this study concludes through empirical analysis that the core of ”friendliness” lies in the unification of high



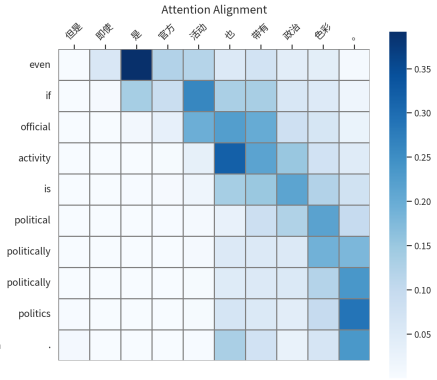
(a) Bi-LTSM on 10k Dataset



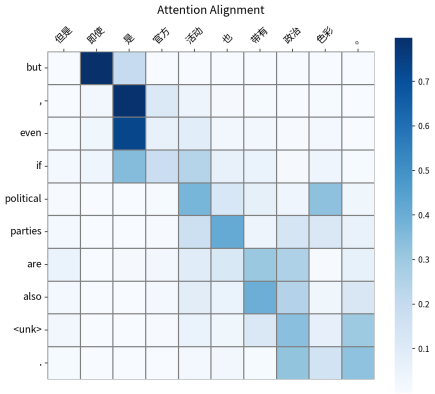
(b) Bi-LTSM on 100k Dataset



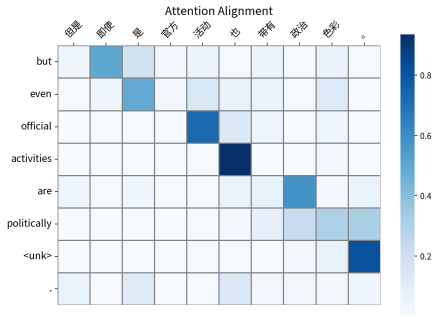
(c) Stacked-LTSM on 10k Dataset



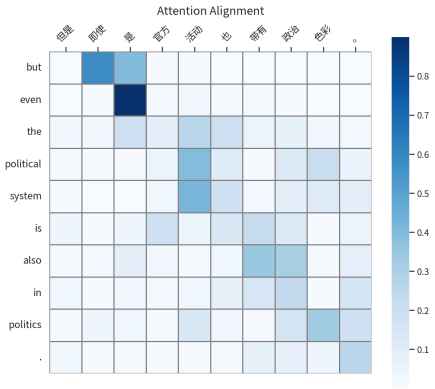
(d) Stacked-LTSM on 100k Dataset



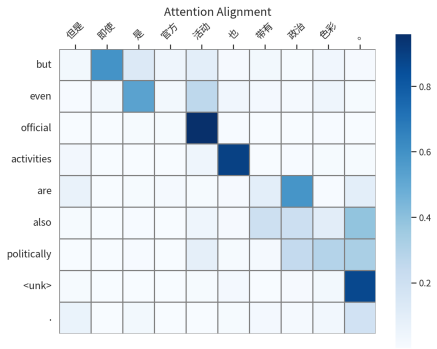
(e) Standard Transformer on 10k Dataset



(f) Standard Transformer on 100k Dataset



(g) Optimal Transformer on 10k Dataset



(h) Optimal Transformer on 100k Dataset

Fig 7: Visualization of Attention Alignment Heatmaps

performance and high interpretability. The experimental results conclusively demonstrate that compared to LSTM architectures constrained by serial computation and fuzzy context attention, the Optimal Transformer, incorporating sparse and grouped-query mechanisms, more accurately mimics the semantic focusing and alignment logic inherent in human language processing. This architectural advantage is reflected not only in significant improvements in translation quality but also in providing a transparent decision-making process for "black box" models through clear diagonal attention matrices, achieving synchronization between computational mechanisms and human cognition. Meanwhile, experiments on dataset scale reveal the trade-off between "exact matching" and "semantic generalization" under limited computational resources, indicating that mere data accumulation cannot replace the inductive bias provided by architectural design in feature extraction; efficient architecture remains the key to resolving underfitting and enhancing generalization capabilities.

Looking forward, to further deepen the human-friendliness of machine translation, future research should focus on deeper semantic interaction and more efficient deployment experiences. On one hand, future work should explore integrating explicit linguistic priors, such as syntactic dependency trees, into attention mechanisms to enhance the logical robustness of models in low-resource or complex sentence scenarios, thereby reducing semantic hallucinations. On the other hand, addressing the inference latency challenges posed by large-scale models, the exploration of lightweight technologies such as model quantization and knowledge distillation will be crucial. These directions will propel machine translation to evolve from a mere text conversion tool into an intelligent linguistic interaction system that is real-time responsive, logically self-consistent, and interpretable.

6 Acknowledgment

I would like to express my sincere gratitude to Prof. Xiaojun Quan for his mentorship in the field of Natural Language Processing. His professional insights and dedicated guidance have provided me with sufficient knowledge in this field and offered great encouragement and help for my future research. I am equally thankful to my fiancée, Ms. Ma Yujie, for her unwavering support and quiet encouragement. Natural Language Processing is a subject I consider myself relatively weak in, but I believe it will yield diverse gains in the future multimodal environment. I hope all classmates and colleagues will continue to steadfastly contribute to scientific development.

References

[1] 屠可伟, 王新宇, 曲彦儒, 等. 动手学自然语言处理 [M]. 第 1 版. 北京: 人民邮电出版社, 2024.

[2] Aston Zhang, Zachary C. Lipton, Mu Li, Alexander J. Smola. 动手学深度学习 PyTorch 版 [M]. Xiaoting He, Rachel Hu. 第 1 版. 北京: 人民邮电出版社, 2023.

[3] Chickering D M. Optimal structure identification with greedy search[J]. Journal of machine learning research, 2002, 3(Nov): 507-554.

[4] Deng Y, Guo D, Guo X, et al. MQA: Answering the question via robotic manipulation[J]. arXiv preprint arXiv:2003.04641, 2020.

[5] Greff K, Srivastava R K, Koutník J, et al. LSTM: A search space odyssey[J]. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.

[6] Kashid S, Kumar K, Saini P, et al. Bi-RNN and Bi-LSTM based text classification for amazon reviews[C]//International conference on deep learning, artificial intelligence and robotics. Cham: Springer International Publishing, 2022: 62-72.

[7] Luo W, Liu W, Gao S. A revisit of sparse coding based anomaly detection in stacked rnn framework[C]//Proceedings of the IEEE international conference on computer vision. 2017: 341-349.

[8] Steinbiss V, Tran B H, Ney H. Improvements in beam search[C]//ICSLP. 1994, 94: 2143-2146.

[9] Tay Y, Bahri D, Yang L, et al. Sparse sinkhorn attention[C]//International conference on machine learning. PMLR, 2020: 9438-9447.

[10] Toomarian N B, Barhen J. Learning a trajectory using adjoint functions and teacher forcing[J]. Neural networks, 1992, 5(3): 473-484.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[12] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.