IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/45AI/2025/4520925/
Vol. 2025 (20), Article ID 4520925

Comparative Analysis of Attention Mechanisms in Neural Text Summarization*

Daniel SIARA and Przemysław CZUBA

Faculty of Cybernetics, Military University of Technology, Warsaw, Poland,

Correspondence should be addressed to: Daniel SIARA, danielsiara111@wp.pl

* Presented at the 45th IBIMA International Conference, 25-26 June 2025, Cordoba, Spain

Abstract

Automatic text summarization is indispensable for navigating today's information overload, yet many organisations cannot afford to train or serve the latest Transformer models. Although additive (Bahdanau) and dot-product (Luong) attention remain the cornerstones of recurrent sequence-to-sequence (Seq2Seq) summarizers, the literature still lacks a controlled, head-to-head comparison of these mechanisms under identical conditions. We therefore built twin LSTM-based Seq2Seq systems that differed only in the attention layer and trained them on an 80 k/20 k article–title split of the Gigaword corpus, using 100-dimensional GloVe embeddings and the Adam optimiser with scheduled-decay teacher forcing, dropout and gradient clipping. Model quality was monitored through validation loss and ROUGE-1/-2/-L F1, complemented by visual inspection of attention heat-maps. The Luong variant converged in fewer epochs, achieved lower validation loss and outperformed the Bahdanau model on all ROUGE metrics, delivering summaries that were consistently more coherent and less repetitive. The additive model remained competitive but required more intensive hyper-parameter tuning and occasionally omitted salient details. Together, these results show that dot-product alignment offers the best accuracy-to-cost trade-off for resource-constrained abstractive summarization and provides a rigorous baseline for future work that augments Seq2Seq designs or benchmarks them against lightweight Transformer architectures.

Keywords: deep neural networks, natural language processing, automatic text summarization, recurrent neural networks, attention mechanism

Introduction

Automatic text summarization has become increasingly significant in an era of information overload, enabling users to obtain the main points of lengthy documents in a compact form. It addresses the challenge of condensing large texts while preserving their key information and meaning. Early summarization approaches relied on manual rules or simple statistical methods, which often struggled with balancing brevity and fidelity to the original content. The advent of deep learning revolutionized this field, as neural networks can learn to generate summaries that capture important content more effectively than heuristic methods. In particular, recurrent neural networks (RNNs) such as LSTM and GRU brought breakthroughs in sequence tasks by handling long-range dependencies in text. These RNN-based sequence-to-sequence (Seq2Seq) models became a foundation for abstractive summarization, capable of producing new sentences that distill the source text's meaning. A critical component that enhanced Seq2Seq models is the attention mechanism, which allows the model to prioritize key segments of the input during decoding. Attention addresses the limitation of vanilla RNNs that may have difficulty capturing all details of very long sequences. By learning to align each generated word with relevant parts of the source text,

Cite this Article as: Daniel SIARA and Przemysław CZUBA, Vol. 2025 (20) "Comparative Analysis of Attention Mechanisms in Neural Text Summarization " Communications of International Proceedings, Vol. 2025 (20), Article ID 4520925, https://doi.org/10.5171/2025.4520925

attention enables more informative and contextually appropriate summaries. Bahdanau et al. (2014) first introduced the additive attention approach in neural machine translation, demonstrating how jointly learning to align and translate improves performance.

Subsequently, Luong et al. (2015) proposed a simplified dot-product attention variant, which computes alignment scores by taking dot products between encoder and decoder states. Both attention types have been applied to text generation tasks and are highly relevant to summarization models that need to decide which parts of an article to emphasize. The goal of this research is to compare the Bahdanau and Luong attention mechanisms in the context of neural abstractive summarization. We aim to evaluate how each mechanism affects the quality of the generated summaries and the efficiency of model training. In particular, we examine differences in automatic evaluation metrics and qualitative aspects of the summaries produced by Seq2Seq RNN models using either attention type. By analyzing these two approaches side by side, we seek to identify which attention strategy yields better summary accuracy and discuss the implications for designing effective summarization systems. The remainder of this paper is organized as follows. In the next section, we review related work on attention mechanisms in Seq2Seq models and prior developments in text summarization. We then describe our methodology, including the Seq2Seq model architecture, the specifics of Bahdanau and Luong attention, and our experimental setup (dataset, training procedure, and evaluation metrics). Next, we present the results, comparing performance metrics and examples of summaries from the two models. Finally, we conclude with a discussion of the key findings and suggest directions for future research.

Related work

Neural attention mechanisms have significantly advanced sequence-to-sequence learning, first gaining prominence in machine translation and later in text summarization. Bahdanau, Cho, and Bengio's work (2014) introduced the concept of additive attention, where the decoder learns to align with different parts of the source sequence at each output step. Their model demonstrated that allowing a dynamic focus on source words greatly improved translation quality. This idea was soon adopted in summarization tasks, since summarization can also be viewed as a sequence transduction problem where the model must select and compress information from the input text. For example, Nallapati et al. (2016) developed one of the early abstractive summarization models using an RNN encoder-decoder with attention. They extended the basic Seq2Seq framework with additional features to better identify salient content, illustrating the importance of attention for generating coherent summaries. Luong at al. (2015) proposed an alternative dot-product attention formulation (often called Luong attention). Instead of using a feed-forward network to compute attention weights as in Bahdanau's method, Luong attention directly computes similarity scores by taking the dot product between encoder hidden states and the decoder's current hidden state. This approach is computationally simpler, and it became widely used due to its efficiency. Both Bahdanau and Luong attention mechanisms have been applied in summarization models and other NLP tasks, sometimes with slight variations or combined with additional techniques (such as coverage mechanisms or pointer networks) to further improve summary completeness and correctness. However, a detailed head-to-head comparison of these two attention types in an identical summarization model setting has been lacking. Most prior research either used one fixed attention method or moved on to more complex architectures without isolating the effect of the attention mechanism itself.

In recent years, the field has seen a shift toward Transformer-based models that rely solely on self-attention (as opposed to RNNs). The Transformer architecture by Vaswani et al. (2017) introduced multi-head self-attention and achieved state-of-the-art results in translation, and subsequent large language models have demonstrated impressive summarization capabilities. Despite this progress, RNN-based Seq2Seq models with attention remain an important baseline and are more feasible to train with limited computational resources. Moreover, analyzing RNN attention mechanisms can provide insights into how different alignment strategies influence the summarization output. This work addresses the gap in directly comparing additive vs. dot-product attention in a controlled summarization experiment. By examining their performance on the same dataset and model architecture, we contribute a clearer understanding of each mechanism's strengths and weaknesses in abstractive text summarization. This comparative analysis complements prior research and can inform the development of efficient summarization systems, as well as guide researchers when choosing an attention mechanism for Seq2Seq tasks.

Methodology

Seq2Seq Model

A recurrent sequence-to-sequence (Seq2Seq) architecture is employed to create abstractive text summaries. This approach is divided into two main components:

- Encoder – processes the input text step by step, generating hidden states that capture semantic information at each time step.

- Decoder – uses the encoder's representations to produce the summary, thereby reflecting the main ideas of the source text.

Either a Long Short-Term Memory (LSTM) or a Gated Recurrent Unit (GRU) network is selected for the recurrent layers. Both variants address the vanishing gradient problem, making it possible to capture long-range dependencies. LSTM introduces input, output, and forget gates that regulate the flow of information, enabling the model to retain or discard specific parts of the input sequence effectively. In contrast, GRU merges certain gates into a more compact structure, reducing the number of trainable parameters while still providing comparable performance. Such flexibility in retaining and updating information is particularly useful when summarizing extensive documents that require a broader temporal context.

To reduce overfitting, dropout is applied during training. This procedure randomly disables a subset of neurons in each layer, preventing the network from developing an excessive reliance on particular weights or features. Additionally, teacher forcing is incorporated: at predetermined intervals, the decoder receives the correct token from the target sequence rather than its own previously generated output. This strategy provides more accurate feedback during training and enhances stability, especially in early epochs. In some cases, a scheduling approach is introduced, gradually lowering the teacher forcing probability over successive epochs to guide the model toward actual inference conditions.

Pretrained GloVe embeddings are also integrated to furnish more informative initial word representations. By loading these embeddings before training, the model leverages previously learned semantic relationships, removing the need to learn them from scratch. This initialization can be especially beneficial when the training dataset is limited, as it accelerates convergence and often leads to improved performance.

Overall, the Seq2Seq design is considered highly adaptable. Model hyperparameters, such as dropout rates or the hidden dimension size, can be tuned to suit specific requirements, and the recurrent backbone can be switched seamlessly between LSTM and GRU. Moreover, various attention mechanisms—discussed later—can be incorporated to further enhance summary coherence. This adaptability, combined with teacher forcing and pretrained embeddings, enables the production of concise yet faithful summaries that accurately reflect the core content of the source text.

Attention Mechanisms

Two attention mechanisms are incorporated to address the issue of relying exclusively on the final encoder hidden state, thereby granting the decoder more direct access to any part of the input sequence. This approach helps capture critical segments of the text at each decoding step, improving summary quality and coherence.

- Bahdanau Attention (Additive)

  Also referred to as additive attention, this method was introduced in the context of neural machine translation by Bahdanau et al. (2014). In this approach, a small feed-forward network is used to compute an energy score that aligns the decoder's hidden state with each encoder hidden state. A typical implementation employs a nonlinear activation function (e.g., tanhf_0\tanh) for this scoring step. The resulting energies are normalized through a softmax function to obtain attention weights, which indicate the relevance of each source position to the current decoding step. By incorporating a learnable nonlinear transformation, Bahdanau attention can capture more nuanced relationships between encoder and decoder states. It effectively pinpoints the most important parts of the input text, allowing the decoder to concentrate on relevant words or phrases. This capability helps mitigate the challenge of losing information in longer sequences, since the decoder is no longer limited to a single vector representation of the entire document. Although the additive operation provides flexibility, it entails additional computation compared to simpler methods, as extra parameters and layer operations are introduced.

- Luong Attention (Dot-Product)

  Often called dot-product attention, this variant was proposed by Luong et al. (2015) to simplify the alignment score calculation. Instead of a separate feed-forward network, the decoder's hidden state is directly multiplied (dot product) by each encoder hidden state, yielding scores that reflect their similarity. As with Bahdanau attention, these scores are normalized using softmax to determine which parts of the input sequence are most pertinent for generating the next output token. The primary advantage of Luong attention lies in reduced computational overhead, since it does not require additional nonlinear layers to compute alignment. This simplicity can lead to faster training and easier implementation, especially in large-scale or resource-constrained environments. However, the direct dot-product approach may be less adaptable than additive methods when dealing with more complex relationships in the data, potentially impacting the granularity with which crucial segments are identified.

Regardless of whether additive or dot-product attention is applied, the essential function remains the same: the decoder dynamically focuses on specific segments of the source text during generation. At each time step, an attention distribution is calculated across the encoder states, highlighting only the relevant parts. This mechanism substantially improves the model's capacity to handle long or information-dense documents, since the decoder can revisit earlier sections of the text as needed to maintain coherence. In abstractive summarization, the choice between Bahdanau and Luong attention can influence the fine-grained detail and clarity of generated summaries. Bahdanau attention typically excels in capturing subtle semantic relationships, whereas Luong attention offers a streamlined, computationally efficient alternative that may be preferable if training resources are limited.

Experiments

Dataset

A subset of the Gigaword corpus is employed as the primary source of data for this study. Gigaword is a large-scale collection of newswire articles paired with concise summaries, making it well-suited for abstractive summarization. Although the complete corpus contains millions of text–summary pairs, only a smaller portion is utilized here to accommodate limited computational resources. Specifically, 80,000 article–summary pairs are assigned for training and 20,000 for validation. This scaled-down set still captures a variety of topics and writing styles, ensuring sufficient diversity to train and evaluate the model effectively.

Preprocessing procedures begin with tokenization at the word level, which facilitates the handling of textual data by splitting it into manageable units. Next, the text is normalized to lowercase, helping reduce inconsistencies that can arise from differences in capitalization. Each record contains two key fields: the article, representing the original text passage, and the title, serving as its corresponding summary. This division aligns naturally with the goal of producing concise summaries from longer source documents.

To enhance semantic understanding during training, pretrained GloVe embeddings are applied to the tokenized data. These embeddings provide a richer lexical representation by incorporating knowledge of word relationships gleaned from large external corpora. By leveraging such representations, the model can better capture essential information and nuances from the source text, ultimately improving the quality of the generated summaries.

Tools and Technology

All experiments were conducted in a cloud-based environment provided by Google Colab, which offered access to an NVIDIA A100 GPU. This powerful hardware setup shortened model training times for the recurrent networks, enabling more extensive experimentation within feasible time limits. The Colab platform also simplified the installation of necessary libraries and supported seamless integration with Google Drive for data and model management.

The primary software foundation for this work was Python, complemented by a range of libraries critical to deep learning and natural language processing tasks. PyTorch served as the core framework, permitting efficient training of Seq2Seq models with GPU-accelerated computation and automatic differentiation. Text preprocessing and vocabulary handling were supported through TorchText, while SpaCy was employed for accurate tokenization. Model evaluation relied on ROUGE metrics, computed using a dedicated library designed for

measuring the overlap between generated and reference summaries. Visualization tools, such as matplotlib, facilitated the examination of attention weight distributions and the monitoring of training dynamics. Additionally, TensorBoard was employed to track the evolution of the loss function and other performance indicators in real time, aiding in the iterative refinement of hyperparameters.

This combination of hardware resources and Python-based technologies ensured that complex RNN architectures with attention could be trained and evaluated efficiently, thereby providing a reliable experimental setup for advancing abstractive summarization research.

Training Setup

A cross-entropy objective is applied to compare the model's predicted tokens with the reference summaries, guiding the training process through backpropagation. The Adam optimizer is selected due to its adaptive learning rate and robust performance in natural language processing tasks. An initial learning rate is set and may be adjusted based on validation feedback to balance stable convergence with sufficient training progress.

Gradient clipping is introduced to mitigate the problem of exploding gradients, which can destabilize training in recurrent neural networks. This strategy enforces an upper limit on gradient magnitudes, preventing excessively large updates to the model's parameters. Dropout is also employed in both the encoder and decoder to reduce overfitting and promote better generalization.

Different training configurations are tested to examine the impact of hyperparameters. For instance, one configuration might train a Luong attention model for 25 epochs with a relatively large batch size (e.g., 128) and an initial teacher forcing probability of 0.95, which gradually decreases to 0.70. Another configuration might allocate 15 epochs to a Bahdanau attention model, utilizing a smaller batch size (e.g., 64) due to its more complex additive mechanism, along with a higher initial teacher forcing probability that decreases more sharply (e.g., from 1.0 to 0.50). Adjusting the dropout rate, the hidden state size, or the learning rate can further refine performance.

Throughout training, intermediate evaluations are carried out on the validation set to monitor progress and detect potential overfitting or underfitting. Whenever the validation loss improves, the current state of the model can be preserved as a checkpoint to ensure that the best-performing parameters are recorded. By iteratively fine-tuning these hyperparameters—number of epochs, learning rate, batch size, dropout rate, and teacher forcing schedule—each model variant (Bahdanau or Luong attention) can be optimized under the available computational constraints.

Evaluation Metrics

Evaluating the quality of generated summaries is crucial for assessing model performance and guiding further refinements. While validation loss provides an initial indication of how well the model fits unseen data, it does not fully capture whether the generated outputs accurately convey the main ideas of the source text. Therefore, this work uses ROUGE (Recall-Oriented Understudy for Gisting Evaluation) as the primary metric suite. ROUGE is widely accepted in natural language processing for summarization tasks, as it measures the overlap between the predicted summaries and reference summaries.

Three variants of ROUGE are considered:

- ROUGE-1: Focuses on matching individual words (unigrams) between the model's output and the reference. High ROUGE-1 scores indicate that the most essential terms are preserved in the generated summary.

- ROUGE-2: Examines pairs of words (bigrams), offering insight into how well the system captures short-word sequences. This metric evaluates not just the presence of keywords, but also their immediate context.

- ROUGE-L: Evaluates the length of the longest common subsequence (LCS) shared by the generated and reference summaries. Unlike simple n-gram counts, ROUGE-L takes into account the global sequence structure, assessing how faithfully the model reproduces the overall flow of ideas.

Each of these metrics yields three values—precision, recall, and an F1-score. Precision represents the fraction of generated n-grams (or subsequences) that are also present in the reference, while recall measures the fraction of reference n-grams the model captures. F1-score balances these two factors, providing a single measure of summary quality. By comparing average ROUGE scores across different models, configurations, or training regimens, it becomes possible to pinpoint strengths and weaknesses in the summarized outputs and to identify which approaches produce the most faithful, coherent summaries.

Results

Table 1: Results of model training experiments

Experiment	Valid loss	Precision			Recall			F1		
		Rouge1	Rouge2	RougeL	Rouge1	Rouge2	RougeL	Rouge1	Rouge2	RougeL
1. Luong	3.7790	0.0698	0.0253	0.0653	0.4206	0.1743	0.3945	0.1177	0.0432	0.1101
2. Luong	3.7435	0.0676	0.0248	0.0631	0.4082	0.1604	0.3822	0.1146	0.0402	0.1071
3. Luong	3.6531	0.0779	0.0311	0.0726	0.4297	0.1777	0.4020	0.1279	0.0477	0.1196
4. Luong	3.9478	0.0572	0.0195	0.0537	0.3555	0.1201	0.3345	0.1062	0.0287	0.0916
5. Bahdanau	4.9288	0.0493	0.0155	0.0464	0.2628	0.1088	0.2686	0.0823	0.0281	0.0775
6. Bahdanau	5.0808	0.0349	0.0079	0.0327	0.2142	0.0586	0.2021	0.0590	0.0136	0.0555
7. Bahdanau	4.9049	0.0438	0.0119	0.0404	0.2623	0.0834	0.2431	0.0742	0.0205	0.0685
8. Bahdanau	5.024	0.0443	0.0120	0.0418	0.2469	0.0750	0.2308	0.0744	0.0203	0.0693

Table 1 presents the key outcomes of the experiments, comparing Bahdanau and Luong attention methods. Each row shows the validation loss and ROUGE-based results for precision, recall, and F1-score across ROUGE-1, ROUGE-2, and ROUGE-L. Each ROUGE metric provides three values:

- Precision – the fraction of n-grams in the generated summary that appear in the reference
- Recall – the fraction of n-grams in the reference that are present in the generated output
- F1 – the harmonic mean of precision and recall, balancing these two factors into one measure of summary quality

As shown in table, the Luong models achieve lower validation losses and generally exhibit higher ROUGE F1-scores than the Bahdanau models. Among the Luong runs, Experiment 3 obtains the strongest performance, showing relatively higher F1 values across all ROUGE metrics. By contrast, the Bahdanau models produce briefer, often less complete summaries, reflected in lower recall and F1 measures. The gap between these two attention mechanisms is consistent with the findings reported in the literature, though the difference in performance may still depend on data size and hyperparameter optimization.

Examples of Summaries

Luong 1 - Excerpt for summary:
"while global stock markets plunged, a major index of commodity prices, on products from oli to cotton, fell to its lowest level in ## years thursday."

Luong 1 -Summary generated by the model:
"oil prices fall to lowest level in # # years"

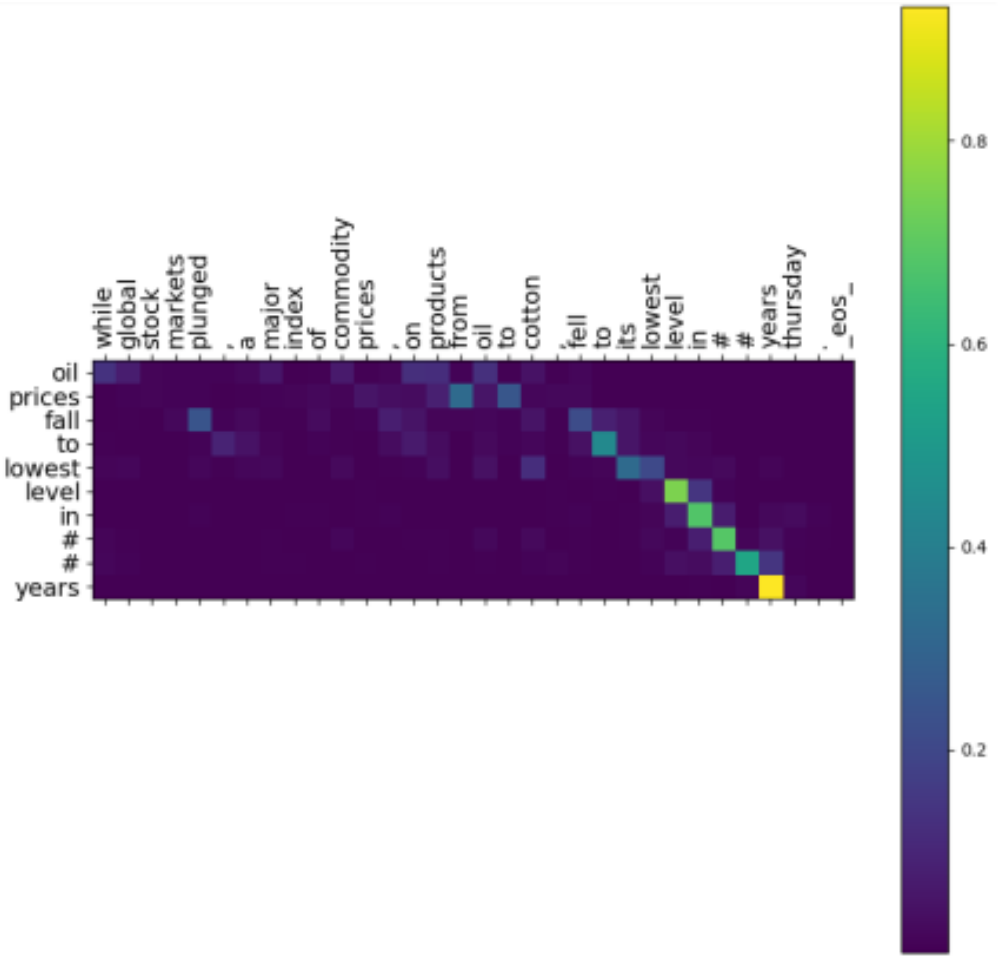

Fig 1. Visualization of the attention matrix Luong 1

Luong 2 - Excerpt for summary:
"Thailand's billionaire prime minister Thaksin Shinawatra is intent on buying an english football club by june and is eyeing league piants Liverpool. The thai national team coach said monday"

Luong 2 - Summary generated by the model:
"thai pm intent buying English football club"

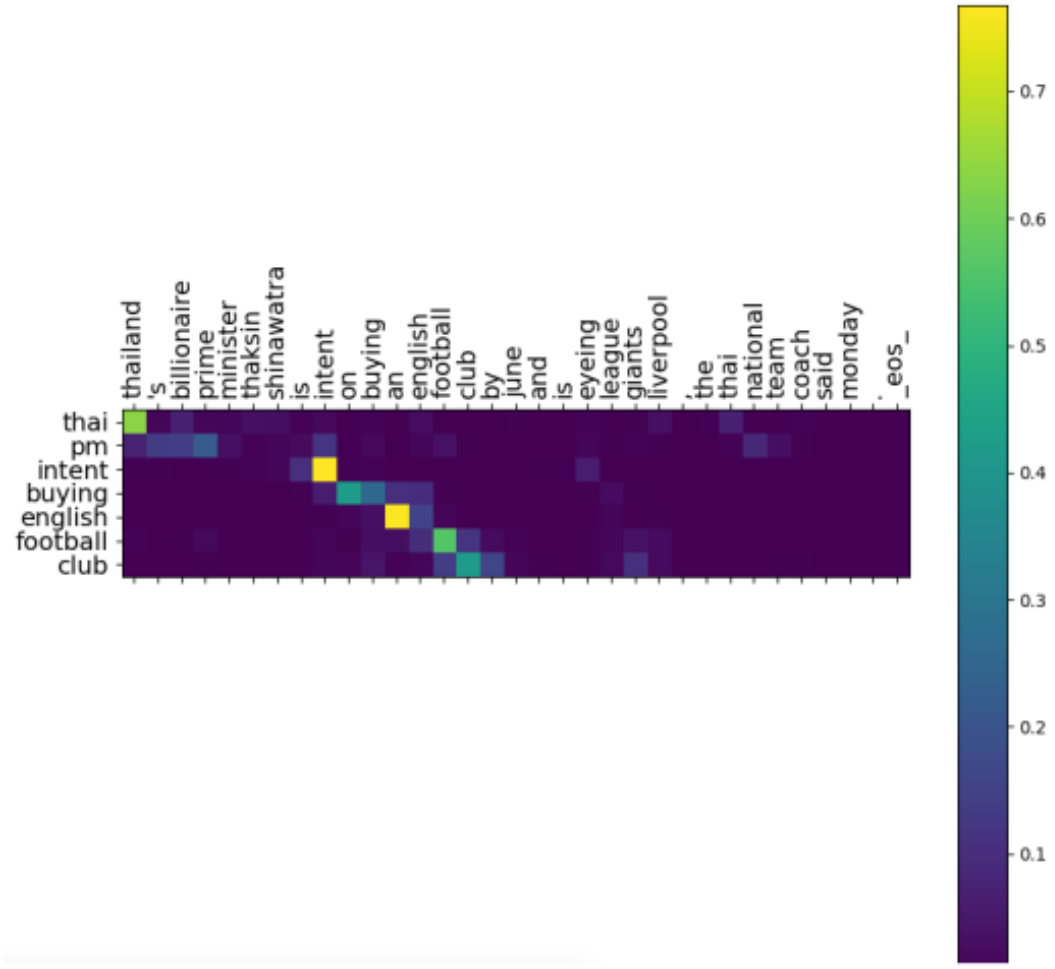

Fig 2. Visualization of the attention matrix Luong 2

Bahdanau 1 - Excerpt for summary:
“Two suicide bombers killed ## people outside a heavily guarded shiite shrine in baghdad on friday , prompting iraq 's prime minister to order an investigation into security shortcomings that allowed the assailants to slip through.”

Bahdanau 1 - Summary generated by the model:
“suicide bombers kill # # in baghdad shrine baghdad”

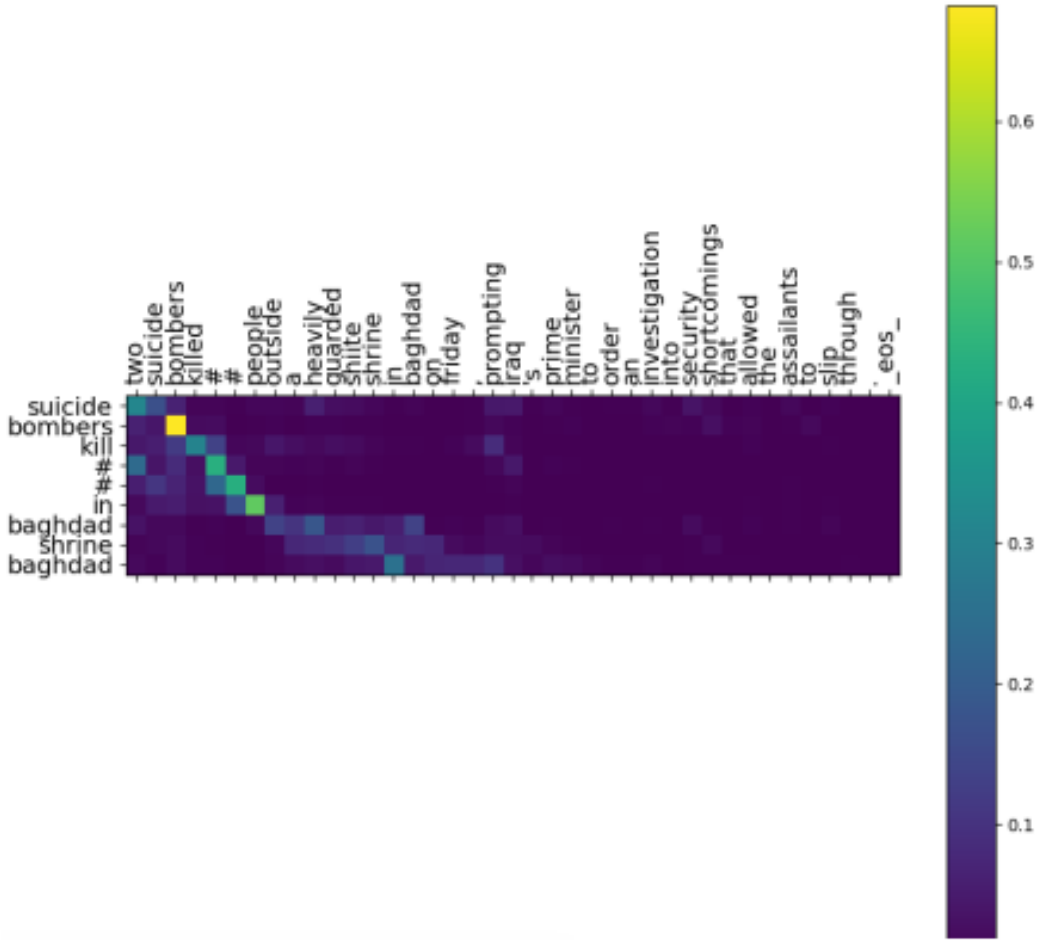

Fig 3. Visualization of the attention matrix Bahdanau 1

Bahdanau 2 - Excerpt for summary:
“the gold price in hong kong rose ## hk dollars at #,### hk dollars a tael monday, according to the bank of china -lrb- hong kong -rrb- , one of the major gold dealers in hong kong.”

Bahdanau 2 - Summary generated by the model:
“gold price in hong kong up”

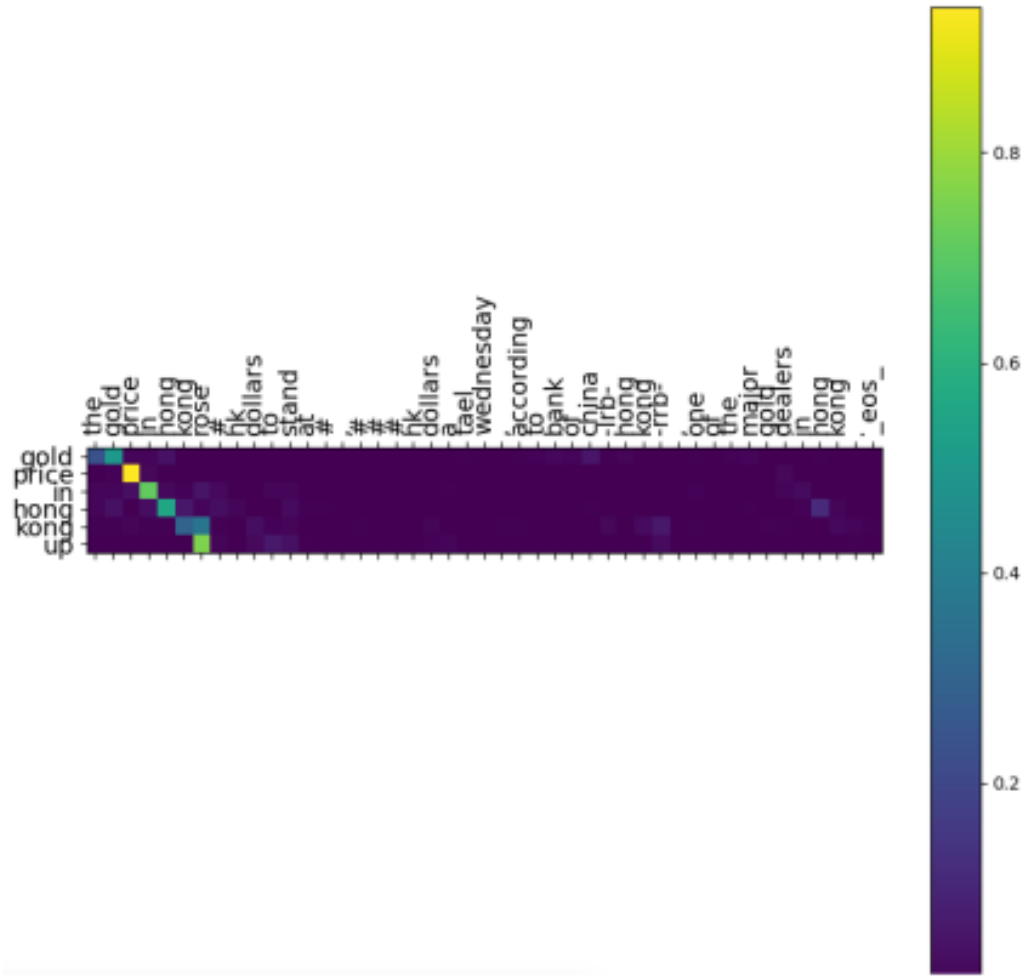

Fig 4. Visualization of the attention matrix Bahdanau 2

Figures 1 and 2 show examples produced by the Luong model. In both instances, the summaries incorporate key terms and phrases from the source text (e.g., “oil prices,” “lowest level,” “thai pm,” “english football club”), indicating a coherent understanding of the main topic. The heatmaps reveal pronounced focus on the most relevant tokens, with attention peaks aligned closely to words conveying essential information.

By contrast, Figures 3 and 4 illustrate Bahdanau-based outputs. While they capture the central theme (e.g., “suicide bombers…in baghdad,” “gold price…hong kong”), these summaries are occasionally shorter and more prone to repetition of certain words. The attention distributions are somewhat broader, with less sharply peaked weights on individual tokens.

Overall, the numeric and qualitative results suggest that Luong attention offers a slight performance advantage in recall-oriented scenarios, likely due to its straightforward dot-product alignment. The Bahdanau mechanism remains competitive, particularly when succinctness is an acceptable trade-off. Statistical significance of the observed differences has not been thoroughly evaluated here, but the consistent trends favoring Luong underscore its practical benefits under the conditions tested.

Conclusions

In this paper, we compared Bahdanau and Luong attention mechanisms in RNN-based sequence-to-sequence models for abstractive text summarization, using a subset of the Gigaword dataset. We aimed to assess their impact on summary quality and model efficiency. Results showed that Luong’s dot-product attention slightly outperformed Bahdanau’s additive attention, achieving higher ROUGE-1, ROUGE-2, and ROUGE-L scores, with summaries that were more detailed and faithful to the source text. Bahdanau’s model, despite its expressive potential, underperformed slightly, producing briefer summaries and requiring more careful tuning. Luong’s

simpler approach trained more easily and generalized better, suggesting it as a practical choice for resource-limited settings. However, Bahdanau might excel with more data or finer alignment needs. Hyperparameter tuning was key for both.

For future work, several avenues emerge from this study. An immediate next step would be to evaluate these attention-based RNN models against the now-dominant Transformer-based models that rely purely on self-attention. Such a comparison would highlight the gap between classic attention in RNNs and the multi-head attention in Transformers for summarization tasks. Additionally, one could experiment with hybrid models or incorporate techniques like coverage mechanisms to further reduce omissions and repetitions, especially for the Bahdanau model to capitalize on its potential.

References

- Bahdanau, D., Cho, K. and Bengio, Y. (2014), 'Neural machine translation by jointly learning to align and translate', arXiv preprint, arXiv:1409.0473, pp. 1–15.
- Luong, M.T., Pham, H. and Manning, C.D. (2015), 'Effective approaches to attention-based neural machine translation', Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1412–1421.
- Nallapati, R., Zhai, F. and Zhou, B. (2016), 'Abstractive text summarization using sequence-to-sequence RNNs and beyond', Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), pp. 280–290.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), 'Attention is all you need', Advances in Neural Information Processing Systems, 30, pp. 5998–6008.