NMT is a neural network that directly models the conditional probability of translating a source sentence to a target sentence. It typically consists of an encoder that computes a representation for the source sentence and a decoder that generates one target word at a time.

NMT is appealing because it requires minimal domain knowledge, is conceptually simple, has a small memory footprint compared to standard MT (as it doesn't need to explicitly store gigantic phrase tables and language models), and NMT decoders are easy to implement. Standard NMT models often use Recurrent Neural Networks (RNNs) such as LSTMs or GRUs for both the encoder and decoder. In non-attentional NMT systems, the source representation is typically used only once to initialize the decoder hidden state.

Attention mechanisms have gained popularity in neural networks for tasks like dynamic control, speech recognition, and image caption generation, allowing models to learn alignments between different modalities. In the context of NMT, attention has been successfully applied to jointly translate and align words. Unlike non-attentional models, attention mechanisms consult a set of source hidden states throughout the entire translation process.

The authors of the paper propose two novel types of attention-based models:

•**Global Attention**: This approach always attends to all source words when deriving the context vector. At each time step in decoding, it takes the target hidden state and derives a variable-length alignment vector by comparing it with each source hidden state. The context vector is then computed as a weighted average over all source hidden states, using the alignment vector as weights. The paper considers different content-based functions (dot, general, concat) and a location-based function to compute alignment scores. This global approach resembles the model by Bahdanau et al. (2015) but is architecturally simpler.

•**Local Attention**: This mechanism chooses to focus on only a small subset of source positions for each target word. This is proposed to address the computational expense of global attention, especially for longer sequences. It draws inspiration from the soft/hard attention trade-off but is designed to be differentiable. The model first predicts a single aligned position (pt) for the current target word. A window centered around pt is then used to compute the context vector as a weighted average of the source hidden states within that window. The local alignment vector is fixed-dimensional. Two variants are explored:

◦

Monotonic Alignment (local-m): Assumes roughly monotonic alignment and sets the aligned position pt equal to the current target time step t.

◦

Predictive Alignment (local-p): Predicts the aligned position pt dynamically using the target hidden state and scales it by the source sentence length. Alignment weights are then modified by a Gaussian distribution centered around the predicted pt to favor positions near it. Local-p is described as differentiable almost everywhere.

The paper also introduces an input-feeding approach where the attentional vectors ($\tilde{h}t$) are concatenated with inputs at the next time steps in the decoder. The goal is to make the model aware of previous alignment decisions and create a very deep network. This approach is proposed to achieve a "coverage" effect, similar to maintaining a coverage set in standard MT.

Experimental Results and Analysis:

• The models were evaluated on the WMT translation tasks between English and German.

• Both global and local attention approaches were found to be effective.

• Attentional models provided significant gains of up to 5.0 BLEU points over non-attentional systems that already included techniques like source reversing and dropout.

• The input-feeding approach provided additional notable gains.

• The local attention model with predictive alignments (local-p) generally performed best among the proposed attention models.

• By ensembling 8 different models, the authors achieved new state-of-the-art (SOTA) results for English to German translation on WMT'14 and WMT'15, outperforming existing best systems (including NMT models and n-gram rerankers) by more than 1.0 BLEU point.

• The unknown word replacement technique yielded further gains, suggesting attention helps in aligning rare words.

• Analysis showed that attentional models are more effective at handling long sentences compared to non-attentional ones, where translation quality does not degrade as sentences become longer.

• Comparison of alignment functions revealed that content-based functions (dot, general) generally performed better than the location-based function, and that dot worked well for global attention while general was better for local attention.

• Evaluating alignment quality using AER showed that learned alignments were comparable to the Berkeley aligner, with local attention models achieving lower AERs than the global one. However, AER and translation scores were not strongly correlated.

• Sample translations highlighted the superiority of attentional models in correctly translating names (e.g., "Miranda Kerr", "Roger Dow") and handling complex phrases like doubly-negated terms.

In conclusion, the paper demonstrates that the proposed simple and effective global and local attention mechanisms significantly improve NMT performance, particularly in handling long sentences and translating specific words like names, leading to state-of-the-art results on English-German translation tasks.