# EvolveMT: an Ensemble MT Engine Improving Itself with Usage Only

**Kamer Ali Yuksel, Ahmet Gunduz, Mohamed Al-Badrashiny,**
**Shreyas Sharma, and Hassan Sawaf**
aiXplain Inc., 16535 Grant Bishop Lane, Los Gatos, CA 95032, US
{kamer, ahmet, mohamed, shreyas, hassan}@aixplain.com

## Abstract

This paper presents EvolveMT for efficiently combining multiple machine translation (MT) engines. The proposed system selects the output from a single engine for each segment by utilizing online learning techniques to predict the most suitable system for every translation request. A neural quality estimation metric supervises the method without requiring reference translations. The online learning capability of this system allows for dynamic adaptation to alterations in the domain or machine translation engines, thereby obviating the necessity for additional training. EvolveMT selects a subset of translation engines to be called based on the source sentence features. The degree of exploration is configurable according to the desired quality-cost trade-off. Results from custom datasets demonstrate that EvolveMT achieves similar translation accuracy at a lower cost than selecting the best translation of each segment from all translations using an MT quality estimator. To our knowledge, EvolveMT is the first meta MT system that adapts itself after deployment to incoming translation requests from the production environment without needing costly retraining on human feedback.

## 1 Introduction

Machine Translation (MT) has experienced substantial progress in recent years, resulting in improving accuracy and more human-like translation output. Despite these advancements, challenges remain, particularly in ensemble modeling. Ensemble models integrate predictions from multiple individual models to achieve a more accurate final output. However, the effective combination of these models is often a complex task that requires thoughtful consideration of factors such as the model architecture, training data, and prediction combination methods. One of the significant challenges in MT ensembling is that the training data used to train the ensemble model, may not be fully representative of the data to be translated later, leading to a mismatch between the model and the data. This paper presents EvolveMT, a method that addresses data drift in ensemble models by continual self-adaptation for optimal performance during usage.

In the subsequent section, we review existing machine translation (MT) quality estimation metrics in the literature, which have been trained on human evaluation or post-editing datasets. In the Approach section, we present a comprehensive explanation of the proposed method. In the Experiments section, we describe our experimental design and provide quantifiable results demonstrating the enhancement resulting from the application of the proposed method, as compared to state-of-the-art quality estimation metrics. Finally, we discuss the obtained results and present our conclusions.

## 2 Related Work

In the WMT20 Metrics Shared Task (Mathur et al., 2020), four reference-free metrics were submitted to evaluate machine translation outputs in the news translation task. These metrics use bilingual mapping of contextual embeddings from language models such as XLM-RoBERTa (Conneau et al., 2019) to assess cross-lingual semantic similarity. However, they often struggle to accurately differentiate between human and machine translations, except for COMET-QE (Rei et al., 2020), the only reference-free metric capable of doing so.

The study by Freitag et al. (2021a) evaluated top MT systems from WMT 2020 using Multidimensional Quality Metrics (MQM) and professional translator annotations. Their results showed a low correlation between crowd worker evaluations and MQM, leading to different rankings and questioning previous conclusions. The study also found that automatic metrics based on pre-trained embeddings can outperform human crowd workers, suggesting that models trained with crowd-sourced

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 5: Industry Track*, pages 341–346
July 10-12, 2023 ©2023 Association for Computational Linguistics

human evaluations may have higher accuracy.

The WMT21 Metrics Shared Task (Freitag et al., 2021b), used MQM expert-based human evaluation to acquire reliable ratings, and evaluate metrics on news and TED talk translations produced by MT systems. Results showed reference-free metrics COMET-QE and OpenKiwi performed well in scoring human translations but not as well with MT outputs, and were strong at segment-level human translation evaluation while competitive with reference-based metrics in system-level evaluation.

REGEMT (Štefánik et al., 2021) is a reference-free metric in WMT21 that uses an ensemble of surface, syntactic, and semantic similarity metrics as input to a regression model. As demonstrated by CushLEPOR (Han et al., 2021), it allows customization, outperforming lexical semantic similarity-based metrics with a higher computational cost.

Onception (Mendonça et al., 2022) used active learning to converge an MT ensemble in a production environment to the best MT with evaluations acquired online.

(Naradowsky et al., 2020) used bandit-learning to adapt MT policies based on simulated user feedback, outperforming the best single MT in mixed-domain settings. A contextual bandit strategy was proposed to make instance-specific decisions, but the system still required a human-in-the-loop (HITL) process.

## 3 Approach

EvolveMT is a quasi MT ensemble technique. In contrast to the traditional multi-system MT approach, which combines outputs from multiple MT systems to enhance translation accuracy and fluency, EvolveMT prioritizes the selection of the most optimal translation from a finite set of MT systems, as we demonstrate in this section. Figure 1 below shows the system architecture of EvolveMT. The system is centered around a multi-class classification model that drives multiple processes to select the best MT model for translation requests.

For each incoming machine translation request, we use SpaCy (Honnibal and Montani, 2017) and Stanza (Qi et al., 2020) frameworks to extract morphological and lexical features. These features include the count of tokens, characters, and the average word length, as well as the frequency of Part-of-Speech labels (such as nouns, verbs, adjectives, etc.), the frequency of Named Entity Recognition labels (including entities such as persons, locations, organizations, etc.), and the frequency of morphological features (e.g. gender and aspect). These features are combined with the 1024-dimensional embedding vector generated by the XLM-RoBERTa encoder of the COMET-QE model and stored alongside the input sentence in the *Ranked Batch Requests Queue*. This queue serves the purpose of prioritizing translation requests that necessitate precedence in processing.

At the outset, requests in the Ranked Batch Requests Queue are ranked based on the order in which they are added. The highest-ranked Machine Translation (MT) request is selected for translation. The *Multi-class MT Classifier* employs the extracted features of the selected MT request to determine the MT systems to be utilized. The classifier prioritizes MT systems with a higher probability of having a higher COMET-QE value. Exploration of additional MT systems becomes more likely only if the probabilities from the classifier's prediction exhibit high entropy. This enables EvolveMT to minimize the cost of exploration when the best MT is predicted with high certainty.

Finally, the selected MT systems are utilized to translate the MT request, and the COMET-QE score is calculated for each translation. The translation with the highest score is chosen and returned in response to the MT request. The Multi-class MT Classifier is then updated online with the best MT system, as determined by the COMET-QE score, serving as a label for the extracted features of the MT request. The online machine learning (ML) functionality of FLAML AutoML framework (Wang et al., 2021) is utilized for online learning. This capability enables the optimization of model hyper-parameters during the iterative course of ML, facilitating continual ML without repetitively hyper-tuning the classifier from scratch.

Subsequently, the classifier is employed to re-rank the Ranked Batch Requests Queue, based on the uncertainty of the classifier, with requests having higher entropy being placed at the top of the queue for prioritized translation. Getting the MT request with the maximum entropy from the queue after each iteration, helps prioritize the most informative sample for the iterative training of the classifier. As the classifier improves its ability to predict the best MT model for MT requests via learning, it reduces the likelihood of exploring other MT(s).

The driving algorithm in EvolveMT, which outlines the primary process of the proposed method, is presented in pseudo-code form in Algorithm 1.

**Algorithm 1** EvolveMT with online active-learning

```
Require: $MTQueue$: list of tuples (source text, features),
  and $MaxMTs$: maximum number of MTs to sample
  while $len(MTQueue)$ do
    $Classifier.rankByUncertainty(MTQueue)$
    $source, feats \leftarrow MTQueue.popMaxEntropyItem()$
    $predMT, classProbs \leftarrow Classifier.predict(feats)$
    $predTrans \leftarrow Translate(source, predMT)$
    $randMTs \leftarrow sampleMTs(classProbs, MaxMTs)$
    $maxEnt \leftarrow normalizedEntropy(classProbs)$
    if $randMTs_0 = predMT$ and $maxEnt < \alpha$ then
      $Classifier.learn(feats, predMT)$
    else
      $sampled \leftarrow Translate(source, randMTs)$
      $randScores \leftarrow CometQE(source, sampled)$
      $predMTScore \leftarrow CometQE(source, predTrans)$
      if $max(randScores) > predMTScore$ then
        $Classifier.learn(feats, randMTs)$
        $IndexOfBestMT \leftarrow randScores.argMax()$
        $predTrans \leftarrow sampled_{IndexOfBestMT}$
      else
        $Classifier.learn(feats, predMT)$
      end if
    end if
    $respondMTRequest((source, predTrans))$
  end while
```

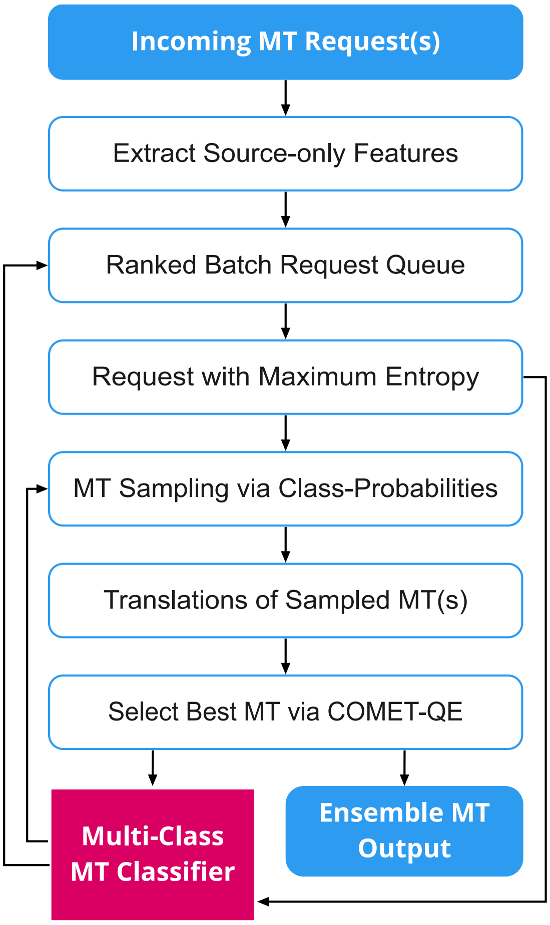
Incoming MT Request(s)
Extract Source-only Features
Ranked Batch Request Queue
Request with Maximum Entropy
MT Sampling via Class-Probabilities
Translations of Sampled MT(s)
Select Best MT via COMET-QE
Multi-Class MT Classifier
Ensemble MT Output

Figure 1: EvolveMT System Architecture

## 4 Experiments

### 4.1 Data

A multi-lingual corpus of 37,500 human-translated sentences in Czech, German, and Russian, along with their corresponding English source-texts, was collected for the OPUS repository (Lison and Tiedemann, 2016; Aulamo and Tiedemann, 2019) using stratified random sampling for each language and dataset. To evaluate EvolveMT, translations for each sentence in the corpus were obtained from one open-source machine translation system (Tiedemann et al., 2022) and five major machine translation service providers in the industry (Google, Azure, AWS, ModernMT, and AppTek).

### 4.2 Setup

Experiments were conducted on a 64-bit Ubuntu 22.04 LTS computer system with an AMD Ryzten 5950X CPU (16 processors, 32 threads) and 64GB of memory. An Nvidia GeForce RTX GPU was used for XLM-RoBERTa embedding extraction from the fine-tuned COMET-QE encoder. The results showed an average 2.88 (±0.06) millisecond response time for the EvolveMT system to return an MT output and update its classifier when the MT request queue contained a single item. Depending on GPU usage, the feature extraction time was (183.43 -244.80) milliseconds. It's worth noting that in a production setting, feature extraction can be performed in parallel for multiple MT requests.

### 4.3 Evaluation criteria and baselines

In this paper, grid search is used to evaluate the impact of two hyperparameters, $maxMTs$ and $\alpha$, on the classifier's performance. $maxMTs$ refers to the maximum number of machine translation systems the classifier can select, and $\alpha$ is the maximum entropy threshold (as described in Algorithm 1). The grid search involves varying $maxMTs$ from 1 to 6 (the maximum number of the MT systems we are using), and $\alpha$ from 0.1 to 1.0 with increments of 0.1. The experimental results are obtained by averaging 100 repetitions to account for the method's inherent stochasticity. For clarity in the results section, we present the results over the $maxMTs$ range while setting $\alpha$ to its optimal value of 0.2, determined through the grid search.

For the evaluation, we adopt the reference-based quality score COMET-DA, as detailed in (Rei et al., 2020), as the evaluation metric for our ensemble output. This choice is motivated by the results of prior research which have demonstrated that COMET-DA exhibits a higher correlation with human evaluation scores compared to other widely used machine translation metrics, such as BLEU and METEOR. The evaluation of EvolveMT is being conducted against the following baselines:

- COMET-QE Ensemble: translation is performed utilizing the six MT systems. The translation with the highest COMET-QE score is selected for each input sentence as the ensemble translation. Then, the COMET-DA score is calculated using selected translations.

- Best MT: involves translating the entire data using all six MT systems. The MT system that produces the highest overall COMET-DA is then selected as the Best MT to employ.

### 4.4 Results

The comparison of COMET-DA scores of the COMET-QE ensemble and Best MT concerning various variants of EvolveMT with varying $MaxMTs$ values are presented in Table 1. The results are depicted for the three language pairs of English-to-Czech, English-to-German, and English-to-Russian. In addition, the average translation cost of each system across the three languages is also documented in the table. The findings indicate that EvolveMT approximates the COMET-QE ensemble's quality while incurring significantly lower costs.

Furthermore, the results in the table reveal that the optimal cost-quality trade-off for EvolveMT varies depending on the target language. Specifically, for all three language pairs, it can be observed that EvolveMT with $MaxMTs = 3$ and $MaxMTs = 4$ provide the best balance between cost and quality when compared to other individual and ensemble MTs. As $MaxMTs$ increases, EvolveMT can achieve higher MT quality by exploring a larger pool of MT systems from which the best translation can be selected. Hence, the $MaxMTs$ parameter can be adjusted to achieve the desired cost-quality trade-off.

Notably, after only a few hundred Machine Translation (MT) requests from the total dataset, the EvolveMT algorithm demonstrates convergence towards an upper limit of its weighted F1-score, which depends on the parameter $maxMTs$. Figure 2 shows the confusion matrix between the outputs of EvolveMT (with $MaxMTs$ = 4) and the COMET-QE ensemble after 100 translations requests. The swift convergence of EvolveMT with a limited number of requests is mainly due to the utilization of XLM-RoBERTa embeddings that have been fine-tuned specifically for the COMET-QE task. This exemplifies the model's effectiveness, as it begins with no prior knowledge, and within a few hundred requests, it can converge and approach the performance of the COMET-QE ensemble. It is crucial to mention that the results presented in Table 1 encompass the "warm up" phase where EvolveMT starts from zero knowledge until full convergence is achieved. If this phase were excluded, the COMET-DA scores of EvolveMT would likely be even higher.

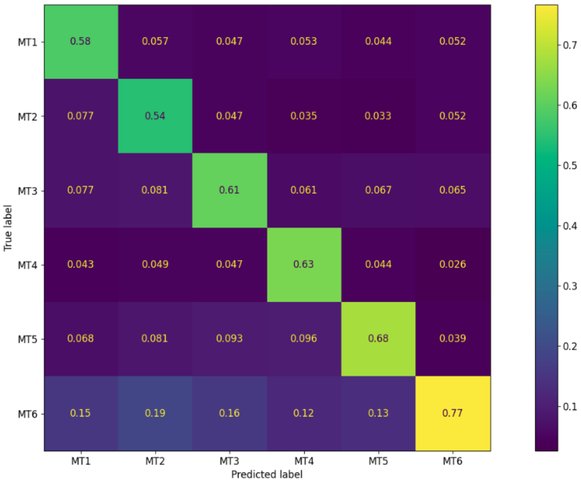

Figure 2: The normalized confusion matrix between EvolveMT ($MaxMTs$ = 4) and COMET-QE after 100 translations requests.

## 5 Discussion

The cost-benefit analysis of EvolveMT highlights the trade-off between run-time efficiency and training expenses. While the run-time cost of EvolveMT may be higher than that of Best MT, it does not require the extensive and time-consuming training process required for traditional MT ensemble methods. This training process involves obtaining translations from all MTs and scoring them using references generated by annotators.

However, the increased run-time cost of EvolveMT is offset by its ability to achieve superior production quality and adapt to changes in the data domain with a minimum amount of overhead. As the data domain changes, traditional MT ensemble techniques require costly retraining to accommodate the new domain, whereas EvolveMT can adapt

|  |  | COMET-DA |  |  |
|---|---|---|---|---|
| **Model** | **Cost ($)** | **English-to-Czech** | **English-to-German** | **English-to-Russian** |
| Best MT (1) | 20.000 | 0.867 | 0.586 | 0.617 |
| COMET-QE (6) | 77.000 | 0.900 | 0.605 | 0.658 |
| EvolveMT (1) | 12.312 | 0.851 | 0.567 | 0.605 |
| EvolveMT (2) | 23.442 | 0.870 | 0.586 | 0.627 |
| EvolveMT (3) | 32.358 | 0.878 | 0.591 | 0.637 |
| EvolveMT (4) | 39.905 | 0.882 | 0.596 | 0.643 |
| EvolveMT (5) | 46.067 | 0.887 | 0.598 | 0.647 |
| EvolveMT (6) | 51.095 | 0.887 | 0.599 | 0.651 |

Table 1: The Cost and COMET-DA comparison of the Best MT system, COMET-QE and EvolveMT ensembles for various $MaxMTs$ parameters (indicated in parentheses). The MT quality increases as COMET-DA scores increase

to changes with a few hundred MT requests.

This versatility and adaptability of EvolveMT make it a robust solution for machine translation tasks that may be subject to data variation, as it can adjust to these changes with minimal effort. The cost-benefit analysis results clearly demonstrate that the increased run-time cost of EvolveMT is outweighed by its high performance and adaptability in the face of changing data domains.

## 6 Limitations

The performance of the EvolveMT system is contingent upon the reliability of the COMET-QE model in providing accurate labels for the MT requests. Utilizing the encoder's embeddings as features necessitates that the COMET-QE model performs effectively on blind MT requests. The batch reranking of MT requests after each learning step may result in a computational bottleneck if the queue size is substantial. To mitigate this issue, an asynchronous re-ranking process could be implemented, whereby the queue is only reorganized once the re-ranking is completed. Additionally, before the re-ranking process, a diverse subset of the queue can be selected based on the XLM-RoBERTa embeddings, which reflect the novelty of the requests relative to previously processed MT requests. The source embeddings from the XLM-RoBERTa model can be cached in parallel during the batch feature extraction process utilizing GPU capabilities, thus facilitating efficient COMET-QE inference. EvolveMT could also be optimized for cost-effectiveness by incorporating the cost of each MT in the ensemble into the algorithm.

## 7 Conclusion and Future Work

This study presents a novel approach called EvolveMT for ensembling machine translation (MT) engines, focusing on minimizing the number of engines required to be queried to achieve optimal quality. To evaluate the efficacy of the proposed method, a series of experiments were conducted, wherein EvolveMT was implemented with varying levels of granularity in terms of the maximum number of engines permitted for each individual MT request. The quantitative results of the experiments indicate that, compared to the traditional method of querying all available MT engines, EvolveMT offers a more cost-effective solution for the ensembling process without compromising the quality of the resulting translations.

EvolveMT presents a unique advantage in terms of cost efficiency compared to COMET-QE Ensemble. This is achieved by utilizing a stochastic exploration approach that selectively queries additional MT engines based on predicted probabilities, which are also employed in an active-learning framework by re-ranking MT requests after each learning step. Furthermore, unlike traditional MT ensemble techniques, EvolveMT can adapt in real-time to changes in customers' translation requests, without incurring the cost of acquiring human references or undergoing costly re-training or fine-tuning.

In conclusion, this paper presents four significant contributions to the field of machine translation: (1) the introduction of the first self-improving MT system that operates without the need for human feedback; (2) the capability of adaptively optimizing the MT ensemble in response to production environment translation requests through online machine-learning; (3) the development of a novel approach for selectively querying MT engines rather than relying on translations from all available engines; and (4) the implementation of an active-learning framework that leverages uncertainties from the ensemble for batch translation.

## 8 Ethics and Impact Statement

EvolveMT is a high-quality machine translation (MT) system for individuals or organizations. It can improve translation accuracy if validated on a specific MT corpus. EvolveMT is trained from scratch for each customer or project, eliminating biases in the algorithm but may still present biases in the quality estimation metric or training dataset. The system is self-adaptable, secure, and protects user privacy by deleting data immediately after translation. EvolveMT eliminates the need for re-training and re-hypertuning, reducing computational costs and being environmentally friendly. The only potential harm is to linguists who perform post-editing as it reduces their dependence on references or evaluations.

## References

Mikko Aulamo and Jörg Tiedemann. 2019. The OPUS resource repository: An open package for creating parallel corpora and machine translation services. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 389–394, Turku, Finland. Linköping University Electronic Press.

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*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. cushlepor: customising hlepor metric using optuna for higher agreement with human judgments or pre-trained language model labse. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1014–1023.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725.

Vânia Mendonça, Ricardo Rei, Luisa Coheur, and Alberto Sardinha. 2022. Onception: Active learning with expert advice for real world machine translation. *arXiv preprint arXiv:2203.04507*.

Jason Naradowsky, Xuan Zhang, and Kevin Duh. 2020. Machine translation system selection from bandit feedback. *arXiv preprint arXiv:2002.09646*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Michal Štefánik, Vít Novotnỳ, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. *arXiv preprint arXiv:2109.07242*.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2022. Democratizing machine translation with opus-mt.

Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. 2021. Flaml: a fast and lightweight automl library. *Proceedings of Machine Learning and Systems*, 3:434–447.