English-to-French Machine Translation on the Europarl Corpus: A Comparative Study of Three Approaches

Maxime Hakawaya **Juan Kenichi Sutan** **Amine Outmani**
Université Paris-Saclay
{maxime.hakawaya-ivanovic, juan-kenichi.sutan, amine.outmani}
@universite-paris-saclay.fr

Abstract

Machine translation (MT) remains a central challenge in natural language processing. In this work, we compare three approaches to English-to-French translation, all trained and evaluated on the Europarl parallel corpus from the OPUS collection. We implement two baselines: (1) a word-for-word model using a positional bilingual dictionary and (2) a retrieval-based model that uses cross-lingual sentence embeddings. We then compare these against (3) our own fine-tuned Seq2Seq transformer, built on top of a pretrained MarianMT model. All models are evaluated on a held-out test set using BLEU score, with both quantitative and qualitative analyses. Our fine-tuned model significantly outperforms both baselines, reaching an average BLEU score of 0.3224 versus 0.0378 and 0.0488 for the word-for-word and embedding baselines. An n-gram breakdown further shows that our model scores well at all n-gram levels, pointing to both accurate word choice and fluent phrasing.

1 Introduction

Machine translation (MT) is one of the oldest problems in artificial intelligence (Hutchins, 2004). Over the decades, the task of automatically translating text between languages has moved from rule-based systems to statistical methods and, more recently, to neural approaches (Stahlberg, 2020).

The Europarl corpus (Koehn, 2005), which contains proceedings of the European Parliament and is distributed as part of the OPUS collection (Tiedemann and Thottingal, 2020), is a widely used benchmark in MT research. It offers large-scale, sentence-aligned parallel text across many European language pairs, which makes it a good fit for training and evaluating translation systems.

In this study, we focus on the English-to-French direction. All three of our models are trained and evaluated on the same Europarl data. We compare two baseline systems against our own fine-tuned model:

1. **Baseline 1 (Word-for-Word)**: A simple model that builds a bilingual dictionary from positionally aligned training pairs.
2. **Baseline 2 (Embedding Retrieval)**: A retrieval model that finds the most semantically similar French sentence from a candidate pool using cross-lingual embeddings.
3. **Our Model (Fine-Tuned Seq2Seq)**: A pretrained MarianMT encoder-decoder transformer that we fine-tune on the Europarl training data, with hyperparameters selected through grid search.

This progression lets us see how each step, from simple lexical lookup to semantic retrieval to learned sequence generation, affects translation quality. We evaluate all models with BLEU score (Papineni et al., 2002) and report both aggregate metrics and per-n-gram breakdowns.

2 Related Work

2.1 Statistical and Neural Machine Translation

MT has gone through several major shifts over the years. Early statistical MT (SMT) systems like phrase-based models (Koehn et al., 2003) relied on word and phrase alignment statistics from parallel corpora. These were eventually replaced by neural machine translation (NMT) approaches, which learn the translation task end-to-end with encoder-decoder architectures (Sutskever et al., 2014; Bahdanau et al., 2015).

The Transformer architecture (Vaswani et al., 2017) was a major breakthrough, allowing for highly parallel training and better handling of long-range dependencies through self-attention. Pre-trained models like MarianMT (Junczys-Dowmunt

et al., 2018) and mBART (Liu et al., 2020) have since made high-quality NMT much more accessible through transfer learning and fine-tuning.

2.2 Cross-Lingual Representations

Cross-lingual sentence embeddings (Reimers and Gurevych, 2020) project sentences from different languages into a shared vector space. Models like paraphrase-multilingual-MiniLM (Wang et al., 2020) make it possible to compare meaning across languages without explicit translation. Although these representations are mainly intended for tasks like semantic similarity and retrieval, they can also be used for translation by returning the closest target-language sentence from a pool of candidates.

2.3 The Europarl Corpus

The Europarl corpus (Koehn, 2005), hosted within the OPUS collection (Tiedemann and Thottingal, 2020), is one of the most commonly used resources in MT research. It contains proceedings of the European Parliament, aligned at the sentence level across up to 21 European languages. Its formal register and high alignment quality make it well suited for benchmarking translation systems. In our project, all three models (both baselines and our fine-tuned model) are trained and evaluated on the same English-French Europarl subset.

3 Methodology

3.1 Data

We use the English-French portion of the Europarl parallel corpus. From the full dataset of roughly 2 million aligned sentence pairs, we load a subset of 100,000 pairs. These are split into training (80%), validation (10%), and test (10%) sets using random sampling with a fixed seed for reproducibility.

3.2 Baseline 1: Word-for-Word Translation

Our first baseline builds a bilingual dictionary from the training data. For each English-French sentence pair, words are aligned based on their relative positions. Concretely, for an English word at position i in a sentence of length n_e, the corresponding French position is estimated as $\lfloor (i/n_e) \cdot n_f \rfloor$, where n_f is the French sentence length. To improve coverage, neighboring positions (± 1) are also considered with reduced weight.

For each English word, the most frequently aligned French word is stored. At inference time, each English word is independently replaced by its dictionary entry, and unknown words are passed through unchanged.

This approach is intentionally simple and serves as a lower bound. It can pick up some word-level correspondences but completely ignores word order, grammar, and context.

3.3 Baseline 2: Cross-Lingual Embedding Retrieval

Our second baseline uses pretrained multilingual sentence embeddings from the `paraphrase-multilingual-MiniLM-L12-v2` model (Wang et al., 2020; Reimers and Gurevych, 2020). During training, we embed all unique French sentences from the training set (up to 50,000 candidates) into a shared multilingual vector space.

At test time, we embed the English input with the same model and return the French candidate with the highest cosine similarity. Figure 1 shows a PCA projection of this shared embedding space. While this approach does capture cross-lingual semantic similarity, it has an inherent limitation: it can only return a sentence that already exists in the candidate pool. If no close match is available, the retrieved sentence may be topically related but quite different from the expected reference.

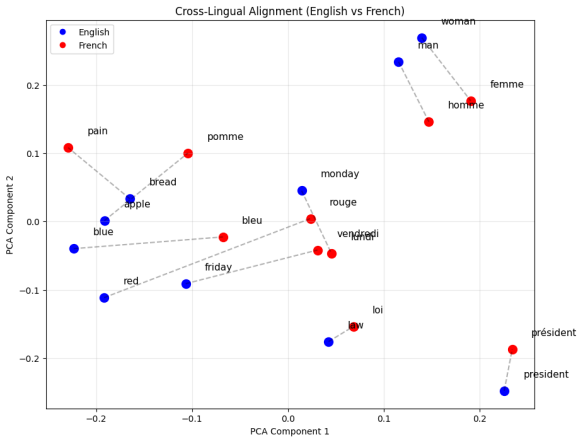

Figure 1: PCA projection of cross-lingual sentence embeddings. English and French translations of the same sentence are mapped to nearby points in the shared space, illustrating the alignment used by Baseline 2.

3.4 Our Model: Fine-Tuned Seq2Seq Transformer

For our main model, we start from the pretrained Helsinki-NLP/opus-mt-en-fr checkpoint (Junczys-Dowmunt et al., 2018), which is a MarianMT

encoder-decoder Transformer already trained on a large collection of English-French parallel data. We then fine-tune this model on the same Europarl training data used by both baselines, selecting hyperparameters through grid search.

It is worth noting that, unlike the two baselines which we built from scratch, this model leverages pretrained weights that already encode substantial translation knowledge. Our contribution here is the fine-tuning and hyperparameter selection on the Europarl domain.

3.4.1 Hyperparameter Selection

We run a grid search over the following hyperparameters:

- Learning rate: $\{2 \times 10^{-5},\ 5 \times 10^{-5}\}$
- Label smoothing: $\{0.0,\ 0.1\}$

Each configuration is trained for 1 epoch with a batch size of 32 and weight decay of 0.01, using mixed-precision (fp16) training on a GPU. We select the best configuration based on validation loss.

The grid search results are shown in Table 1.

LR	Label Smooth.	Val Loss
2e-5	0.0	1.1218
2e-5	0.1	2.4475
5e-5	0.0	**1.1207**
5e-5	0.1	2.4483

Table 1: Grid search results. Best: lr=5e-5, label smoothing= 0.0.

The best configuration (learning rate 5×10^{-5}, no label smoothing) is used to train the final model for 3 epochs. Interestingly, label smoothing consistently hurt performance, roughly doubling the validation loss.

3.4.2 Training Details

The final model is trained with: batch size 16, learning rate 5×10^{-5}, weight decay 0.01, 500 warmup steps, max sequence length 128, and fp16 mixed precision. The best checkpoint by validation loss is saved for evaluation. Figure 2 shows the loss curves, which indicate stable convergence with no signs of overfitting.

4 Evaluation

4.1 Metrics

We evaluate all three models using sentence-level BLEU score (Papineni et al., 2002) with smoothing

Figure 2: Training and validation loss over training steps. The training loss (blue) decreases steadily while the validation loss (red) converges, indicating stable fine-tuning.

(method 1 from `nltk`). BLEU measures n-gram overlap between predicted and reference translations, and is a standard automatic metric for translation quality. We also report exact match rate.

All models are evaluated on the same 10,000 test examples from the Europarl data.

4.2 Overall Results

Metric	BL 1	BL 2	Ours
Avg. BLEU	0.0378	0.0488	**0.3224**
Exact Match	0.0005	0.0071	0.0351

Table 2: Comparison of both baselines (BL 1, BL 2) and our model on 10,000 test examples.

Our fine-tuned model significantly outperforms both baselines, with a BLEU score of 0.3224. This is roughly 8.5× higher than Baseline 1 and 6.6× higher than Baseline 2. The gap between the two baselines themselves is small (0.0378 vs. 0.0488), which suggests that semantic retrieval on its own is not enough for good translation.

4.3 N-gram Level Analysis

To better understand where each model succeeds or fails, we break down the BLEU score by n-gram level (see also Figure 3).

Model	B-1	B-2	B-3	B-4
BL 1	0.3233	0.0696	0.0211	0.0106
BL 2	0.2124	0.0650	0.0375	0.0263
Ours	**0.5857**	**0.3884**	**0.2830**	**0.2110**

Table 3: N-gram BLEU breakdown (B-n = BLEU-n).

Baseline 1 gets a reasonable BLEU-1 of 0.3233, which means it does produce many correct individ-

ual words. But its scores drop off quickly at higher n-gram levels (BLEU-4 = 0.0106), confirming that word-for-word translation gives no coherent phrasing.

Baseline 2 actually has a *lower* BLEU-1 than Baseline 1 (0.2124 vs. 0.3233). This is because it retrieves whole sentences that may be semantically close but use different vocabulary. It does slightly beat Baseline 1 at BLEU-3 and BLEU-4, since retrieved sentences can occasionally overlap structurally with the reference.

Our model keeps high scores at all n-gram levels (BLEU-1 = 0.5857, BLEU-4 = 0.2110), showing both good word choice and fluent multi-word output.

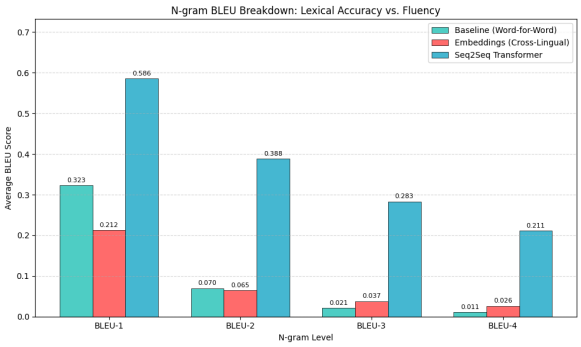

Figure 3: N-gram BLEU breakdown across all three models. Our model (blue) keeps high scores at all n-gram levels, while Baseline 1 drops sharply beyond unigrams and Baseline 2 remains low throughout.

4.4 Qualitative Analysis

We also look at specific translation examples to get a sense of the qualitative differences between models.

Example: *"There are two political reasons why our committee and myself as its chairman must speak out about this issue."*

Reference: *"Une double considération politique conduit notre commission, ainsi que moi-même, en tant que Président de cette commission, à élever la voix dans cette affaire."*

Baseline 1: *"Il nous deux politique raisons pourquoi notre commission et moi comme de président doit de de de ce question"*

Baseline 2: *"Ce sujet a soulevé un débat au sein de la commission."*

Ours: *"Il y a deux raisons politiques pour lesquelles notre commission et moi-même, en tant que présidente, devons parler de cette question."*

Baseline 1 outputs a jumble of individually correct French words with no grammatical structure. Baseline 2 picks a topically related but completely different sentence from the pool. Our model produces a fluent and accurate translation that closely matches the reference. Figure 4 shows the cross-attention weights for a sample translation, giving some insight into how the model aligns source and target tokens.

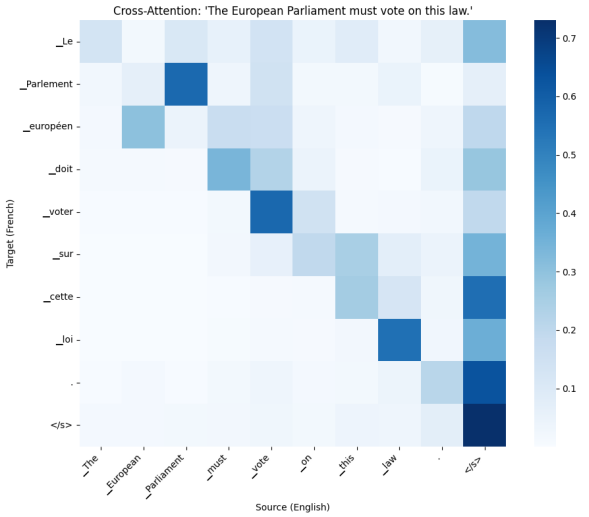

Figure 4: Cross-attention heatmap for our Seq2Seq model. The weights show how each generated French token attends to specific English source tokens, reflecting learned word alignment.

5 Discussion

The results clearly show the expected hierarchy: word-level lookup performs worst, semantic retrieval is slightly better, and the fine-tuned Seq2Seq model is far ahead of both. A few observations are worth highlighting.

Baseline 1 limitations. The word-for-word model cannot capture word order, morphological agreement, or context. Its decent BLEU-1 shows it can get individual words right, but the lack of any multi-word coherence makes its output essentially unusable.

Baseline 2 limitations. The embedding retrieval model can only return sentences from its candidate pool, so it can never produce a genuinely new translation. If the pool does not contain a close match, the output will be off. Its lower BLEU-1 compared to Baseline 1 also shows that sentence-level semantic similarity does not guarantee lexical overlap, which is what BLEU actually measures.

Benefits of fine-tuning. Our model benefits from the pretrained MarianMT weights, which already encode a lot of translation knowledge from large-scale training. Fine-tuning on Europarl then

adapts it to this specific domain. The key advantage over the baselines is that our model actually *learns to generate* translations rather than relying on lookup or retrieval. We also found that label smoothing hurt performance in this setting, possibly because the Europarl domain is fairly homogeneous and the pretrained model already provides enough regularization on its own.

Sentence length effects. As shown in Figure 5, all models perform worse on longer sentences. Our model is the most robust, holding up relatively well even for sentences of 40–50+ words, while Baseline 1 degrades the fastest.

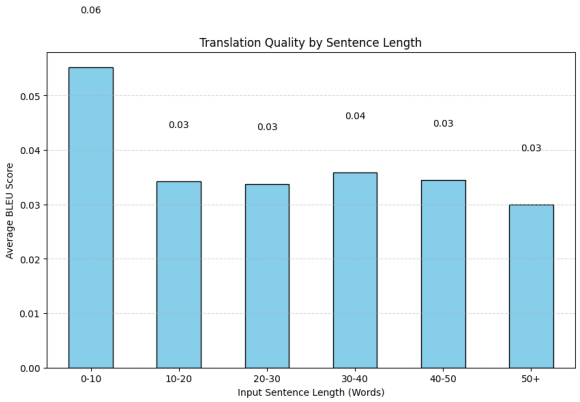

Figure 5: Average BLEU score by input sentence length. All models degrade on longer sentences, but our Seq2Seq model is the most robust.

6 Conclusion

In this project, we compared three English-to-French machine translation approaches on the Europarl corpus: a word-for-word baseline, a cross-lingual embedding retrieval baseline, and a fine-tuned Seq2Seq transformer built on pretrained MarianMT weights. All three models were trained and evaluated on the same data. Our fine-tuned model achieved a BLEU score of 0.3224, far outperforming the two baselines (0.0378 and 0.0488).

The progression from lexical to semantic to generative models gives a clear picture of how different levels of linguistic representation contribute to translation quality. In future work, it would be interesting to try larger training subsets, a broader hyperparameter search, or comparisons with newer model architectures.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

W. John Hutchins. 2004. The Georgetown-IBM experiment demonstrated in January 1954. In Robert E. Frederking and Kathryn B. Taylor, editors, *Machine Translation: From Real Users to Research (AMTA 2004)*, pages 102–114. Springer.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL)*, pages 127–133.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4512–4525.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.