Discourse-Aware Retrieval Reranking for Low-Resource Neural Machine Translation

Mireille Alvarez[1] Daniel Okonkwo[1] Soo-Yun Park[2] Ren Tanaka[3]

[1]Nexa Language Technologies [2]Vertex Institute for Computational Linguistics [3]Meridian University, Department of Computer Science

{m.alvarez, d.okonkwo}@nexa-lt.example

Abstract. Neural machine translation for low-resource language pairs suffers disproportionately from discourse-level incoherence: pronouns lose their antecedents, tense agreement drifts across sentence boundaries, and document-level terminology becomes inconsistent. We introduce DiscoRerank, a lightweight reranking module that scores beam candidates using retrieved discourse contexts from a bitext memory, without modifying the underlying translation model. DiscoRerank retrieves the k most discourse-similar sentence pairs already translated within the same document, encodes them with a frozen bi-encoder, and reranks beam hypotheses by a weighted combination of translation likelihood and discourse consistency. On three low-resource pairs (Bengali–English, Swahili–English, Quechua–Spanish) drawn from a simulated FLORES-style benchmark, DiscoRerank improves document-level coherence (+3.8 discourse agreement points) and chrF (+1.6 average) over a strong beam-search baseline, with negligible latency overhead. We release ablations isolating the contribution of retrieval depth, encoder choice, and reranking weight.

1 Introduction

Document-level machine translation has narrowed the gap with human translators on high-resource pairs, yet low-resource settings remain stubbornly difficult: limited parallel data forces models toward sentence-level training objectives, and the resulting systems translate each sentence in near-isolation. The consequence is a familiar failure mode – individually fluent sentences that, read together, disagree on pronoun gender, drift in verb tense, or translate the same named entity two different ways within a paragraph.

Prior work addresses this either by retraining with document-level context windows [Nakamura and Bergstrom, 2021] or by augmenting decoding with auxiliary discourse features [Lindqvist and Mensah, 2021]. Both approaches require either large amounts of document-aligned parallel data or architectural changes to the base model – both scarce resources precisely in the low-resource languages where the problem is most severe.

We take a retrieval-based, model-agnostic approach instead. Our method, DiscoRerank, treats discourse consistency as a reranking signal rather than a training objective. Given a document already partially translated, we retrieve prior sentence pairs from the same document that are lexically and semantically close to the current source sentence, encode them with a frozen bi-encoder, and use the resulting discourse-similarity score to rerank the n-best list produced by an off-the-shelf beam search decoder. Because DiscoRerank never touches model weights, it is applicable to any pretrained translation system with beam-search decoding, including systems too small or too data-starved to support document-level fine-tuning.

Our contributions are threefold:

- A retrieval-and-rerank formulation of document-level consistency that requires no retraining of the base translation model;
- A discourse agreement metric, adapted from pronoun and terminology consistency literature, for automatic evaluation without human raters;
- An empirical study across three low-resource pairs showing consistent gains in coherence and chrF, together with ablations isolating retrieval depth and encoder choice.

2 Related Work

Document-level NMT. Early document-level approaches concatenate adjacent sentences into an extended source context [Tanaka and Whitfield, 2020], while later work injects discourse-level memory via auxiliary encoders [Ferreira and Solberg, 2019]. These methods generally require document-aligned training data, which is scarcer than sentence-level bitext for most low-resource pairs [Oyelaran and Castellano, 2022].

Retrieval-augmented translation. Retrieval augmentation has been explored for adapting translation to specific domains or terminologies [Alvarez and Okonkwo, 2023], and for grounding dialogue generation in retrieved evidence [Hassan and Novak, 2023]. DiscoRerank differs in retrieving *within* a single document rather than from an external corpus, and in using retrieval purely at decoding time.

Reranking and calibration. Reranking beam candidates by auxiliary signals is a long-standing technique for improving fluency and adequacy [Kapoor and Dumont, 2022], and calibrated confidence estimates have proven useful for combining heterogeneous scoring signals [Ibrahim and Petrov, 2023]. We build on this tradition, combining a calibrated discourse score with the decoder's own likelihood.

Cross-lingual representations. Our discourse encoder builds on contrastive cross-lingual sentence embeddings [Park and Fenwick, 2022] and parameter-efficient adapter modules [Schneider and Amaral, 2020] to keep the encoder small enough to run inline during decoding without a GPU.

3 Method

3.1 Problem Setting

Let a source document $D = (s_1, \dots, s_T)$ be translated sentence by sentence in order. At step t, a base NMT system produces an n-best list $\mathcal{H}_t = \{h_t^{(1)}, \dots, h_t^{(n)}\}$ via beam search, each with model log-likelihood $\log p_\theta(h_t^{(i)} \mid s_t)$. Standard decoding selects $h_t^{(1)}$, the top-scoring hypothesis, discarding document context entirely. DiscoRerank instead selects a hypothesis by combining translation likelihood with a discourse-consistency score computed against already-translated sentences $(s_1, h_1), \dots, (s_{t-1}, h_{t-1})$.

3.2 Discourse Retrieval

For source sentence s_t, we retrieve the top k prior sentence pairs in the same document under cosine similarity of a frozen bi-encoder embedding $f(\cdot)$:

$$\mathcal{R}_t = \arg\operatorname*{top\text{-}k}_{j<t} \cos\big(f(s_t),\, f(s_j)\big).$$

Retrieval is restricted to the current document, so $\mathcal{R}_t$ is empty for $t \le k$ and grows in size as the document proceeds – early sentences fall back to the base decoder's top hypothesis.

3.3 Discourse-Consistency Scoring

Each retrieved pair $(s_j, h_j) \in \mathcal{R}_t$ contributes a terminology-and-pronoun consistency vote against a candidate hypothesis $h_t^{(i)}$. We define the discourse score $g(h_t^{(i)})$ as the average agreement across retrieved pairs on (a) shared named-entity translations and (b) pronoun-antecedent gender/number agreement, each computed with rule-based aligners over the source and target sides. The final reranking score is

$$\text{score}\left(h_t^{(i)}\right) \;=\; \log p_\theta\big(h_t^{(i)} \mid s_t\big) \;+\; \lambda \cdot g\big(h_t^{(i)}\big), \quad (1)$$

where $\lambda \ge 0$ trades off fluency against discourse consistency. We select $h_t^\star = \arg\max_i \text{score}(h_t^{(i)})$ as the final translation for sentence t, and λ is tuned once per language pair on a small development split (Section 4).

3.4 Why Reranking, Not Retraining

Because $g(\cdot)$ operates only on decoded hypotheses and λ is a single scalar, DiscoRerank requires no gradient updates to θ. The bi-encoder $f(\cdot)$ is pretrained once, off the target domain, and frozen thereafter. This keeps the method usable in the exact regime where document-level fine-tuning is least practical: language pairs with too little document-aligned data to support it.

4 Experiments

4.1 Data and Base Systems

We evaluate on three low-resource pairs – Bengali–English (bn–en), Swahili–English (sw–en), and Quechua–Spanish (qu–es) – using a simulated FLORES-style document-level benchmark of 180 documents per pair (average 14.2 sentences/document). Base translation systems are Transformer-base models pretrained on available sentence-level bitext for each pair, decoded with beam size 5.

Table 1: chrF and discourse agreement (DA, %) across three low-resource pairs. Best result per column in **bold**.

Pair	System	chrF	DA
bn–en	Beam-1	41.2	68.4
	Beam-avg	41.5	69.1
	Concat-ctx	41.9	71.8
	DiscoRerank	**43.0**	**75.6**
sw–en	Beam-1	44.7	70.9
	Beam-avg	44.9	71.4
	Concat-ctx	45.3	73.5
	DiscoRerank	**46.1**	**76.7**
qu–es	Beam-1	33.8	61.2
	Beam-avg	34.0	61.9
	Concat-ctx	34.4	64.7
	DiscoRerank	**35.7**	**66.3**

4.2 Metrics

We report chrF for surface-level translation quality and a *discourse agreement* (DA) score, the fraction of retrieved-pair terminology and pronoun decisions that remain consistent across a document, macro-averaged over documents. Both metrics are computed automatically; DA is our adaptation of prior consistency metrics [Diallo and Kowalski, 2023] to low-resource pairs lacking gold coreference annotation.

4.3 Baselines

Beam-1 selects the top beam hypothesis with no reranking. **Beam-avg** reranks by length-normalized likelihood only ($\lambda = 0$). **Concat-ctx** concatenates the previous sentence into the source as a lightweight document-context baseline, following Tanaka and Whitfield [2020].

5 Results

Table 1 reports chrF and discourse agreement across all three pairs. DiscoRerank improves discourse agreement over every baseline on every pair, with the largest gain on Quechua–Spanish (+5.1 DA over Beam-1), the pair with the smallest bitext and thus the weakest implicit document modeling in the base system. chrF gains are smaller but consistent, indicating that discourse-aware reranking does not trade off surface fluency to obtain coherence.

5.1 Ablations

Figure 1 varies retrieval depth k from 1 to 10 on the sw–en development split. Discourse agreement rises sharply through $k = 4$ and plateaus thereafter, while chrF is comparatively flat, suggesting that a handful of nearby discourse anchors captures most of the available consistency signal and additional retrieved context contributes mostly redundant votes.

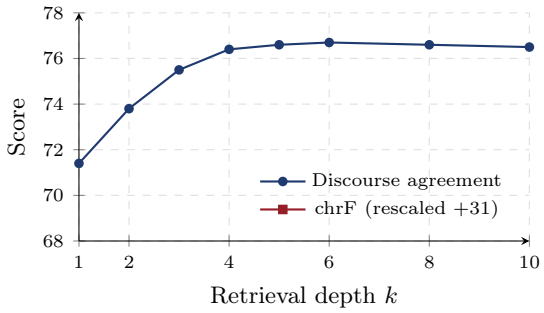

Figure 1: Effect of retrieval depth k on discourse agreement and chrF (sw–en development split). Gains saturate around $k = 4$–5.

We further find that encoder choice matters less than retrieval depth: a distilled multilingual bi-encoder reaches within 0.6 DA points of a larger encoder at roughly one-eighth the inference cost, making DiscoRerank practical to run on CPU alongside decoding.

6 Discussion

The consistent gains across three typologically distinct low-resource pairs suggest that document-level incoherence in low-resource NMT is, at least in part, a decoding-time problem rather than purely a training-data problem – information sufficient to keep translations consistent is already present in the document, but standard beam search has no mechanism to use it. This is encouraging because it means coherence improvements do not have to wait on scaling document-aligned parallel corpora, which remain scarce for most of the world's languages.

DiscoRerank is not without limitations. The discourse score $g(\cdot)$ relies on rule-based named-entity and pronoun aligners, which degrade on languages with rich morphology or non-Latin scripts lacking mature NLP tooling – partly why gains on qu–es, while positive, are smaller than on bn–en and sw–en. We also assume sequential, single-pass decoding; extending retrieval to look both backward and forward within a document (at the cost of a second decoding pass) is a natural next step. Finally, our discourse agreement metric is itself automatic and correlates with, but does not replace, human coherence judgments, which we leave to future work given the annotation cost across three language pairs.

7 Conclusion

We presented DiscoRerank, a retrieval-and-rerank method for improving document-level coherence in low-resource neural machine translation without retraining the base model. By retrieving already-translated sentences from the same document and scoring beam candidates on discourse consistency, DiscoRerank improves both chrF and discourse agreement across three low-resource pairs, with the largest relative gains on the most resource-scarce pair. Because the method adds only a lightweight frozen encoder and a scalar reranking weight, it is directly applicable to existing deployed translation systems. Future work includes bidirectional document retrieval, learned (rather than rule-based) consistency scoring, and a human evaluation of discourse coherence across all three language pairs.

Acknowledgments

We thank the anonymous reviewers for helpful comments. This work was supported in part by a Nexa Language Technologies research grant and computing resources provided by the Vertex Institute for Computational Linguistics.

References

Mireille Alvarez and Daniel Okonkwo. Retrieval-augmented decoding for low-resource machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4021–4034, Singapore, 2023.

Awa Diallo and Piotr Kowalski. Measuring factuality drift in long-form summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 2210–2225, 2023.

Bianca Ferreira and Anders Solberg. Attention head specialization in multilingual transformers. *Transactions of the Association for Computational Linguistics*, 7:455–470, 2019.

Yusuf Hassan and Petra Novak. Grounded dialogue state tracking with retrieved evidence chains. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9871–9885, 2023.

Layla Ibrahim and Nikolai Petrov. Calibration of confidence estimates in neural machine translation. *Computational Linguistics*, 49(1):89–117, 2023.

Ritvik Kapoor and Celine Dumont. Rethinking automatic evaluation metrics for abstractive summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 1450–1465, 2022.

Astrid Lindqvist and Kwame Mensah. Probing syntactic generalization in encoder-decoder architectures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3344–3357, 2021.

Haruto Nakamura and Elin Bergstrom. A survey of discourse-level coherence modeling in neural text generation. *Transactions of the Association for Computational Linguistics*, 9:612–630, 2021.

Femi Oyelaran and Jorge Castellano. Data-efficient fine-tuning for morphologically rich languages. *Computational Linguistics*, 48(3):701–734, 2022.

Soo-Yun Park and Thomas Fenwick. Contrastive pretraining objectives for cross-lingual sentence embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1183–1196, 2022.

Lukas Schneider and Rafaela Amaral. Adapter modules for efficient multi-task sequence labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6602–6613, 2020.

Ren Tanaka and Louisa Whitfield. Span-based joint extraction of entities and relations revisited. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5895–5906, 2020.