Filtering Projected Senses for Bengali: A2 Translation, Target-Language WSD, Dictionary Evidence, and LLM Verification on SE13*

Mohammad Shahriar Hossain
University of Alberta
mhossai6@ualberta.ca

Ekamjot Singh
University of Alberta
ekamjot2@ualberta.ca

Abstract

This paper studies how to *filter* noisy cross-lingual sense projections from English to Bengali on the SemEval 2013 (SE13) subset of XL-WSD. We first obtain an alternative EN→BN translation (A2) using NLLB-200 and compare it with the A1 baseline using COMETKiwi and a small manual evaluation. We then re-project English BabelNet senses onto Bengali via word alignment and evaluate the projected tags with three complementary signals: (i) target-language WSD (Babelfy), (ii) bilingual dictionary evidence (word2word), and (iii) an LLM validator (Gemma-3-4B-Instruct) prompted with bilingual context and glosses. A2 slightly outperforms A1 (0.8488 vs. 0.8189 COMETKiwi); dictionary checks offer broad coverage and useful negative evidence; target-side WSD appears precise but sparse; and the LLM adds flexible context reasoning with some instability. We discuss trade-offs and scaling implications for larger pipelines.

1 Introduction

Projecting English senses to a target language is attractive for bootstrapping sense-annotated resources, but it compounds noise from translation, alignment, and polysemy. We focus on *filtering* projected Bengali (BN) senses after re-running the pipeline with a stronger A2 translation. Three lightweight signals—target-language WSD, dictionary evidence, and an LLM verifier—are compared and analyzed for coverage, reliability, and cost.

Dataset snapshot. On SE13, the English side contains **1,644** gold sense tokens. After A2 translation, alignment, and projection, the Bengali side contains **991** projected sense tokens. Concretely, we quantify how modest MT gains propagate into downstream sense projection quality and systematically compare three post-hoc filters for coverage, precision, and cost. The results point to a simple, scalable recipe: pair a dictionary-first filter with selective WSD/LLM checks for high-value disagreements.

*Code: https://github.com/hossain-shahriar/CMPUT497-A2-Filtering

2 Translation Comparison (A1 vs. A2)

We compare A1 (Google Translate) with an alternative A2 system (NLLB-200).

COMETKiwi (system-level): A1 = 0.8189 vs. A2 = **0.8488**.

A targeted manual check on ten difficult sentences agreed with COMETKiwi. Typical A2 advantages included more natural named-entity renderings and consistent Bengali numerals (e.g., *1990* → ১৯৯০). For example, "U.N." is transliterated as ইউএন by A1 but lexicalized as জাতিসংঘ by A2, which is preferable for Bengali text. Given both automatic and manual evidence, we used A2 for downstream alignment and sense projection.

3 Projected Senses and Filtering Signals

After projection, we subjected BN tokens with candidate senses to three post-hoc signals.

3.1 Target-Language WSD

We applied a Bengali WSD system (Babelfy) to the A2 translations. It produced BN synsets for **612** of **6,822** total BN tokens (low coverage). Among the **991** projected-sense tokens, only **59** matched WSD; **932** disagreed.

Observations. Where present, WSD often corrected mis-projections arising from misaligned English tokens or wrong sense choices; however, sparse coverage limited impact on recall. Illustrative cases:

- অনিয়মিত: WSD → `bn:01796836n` vs. projection → `bn:00066603n`. The WSD sense fits context; the projection reflects upstream alignment error.

- আচরণ: WSD → `bn:00009656n` while projections included `bn:00030015n`/`bn:00009654n`. Dictionary/gloss checks and manual inspection favor the WSD reading.

3.2 Bilingual Dictionary Evidence

We next checked whether each aligned EN–BN pair is attested in a bilingual lexicon (word2word). Of the **991** projected-sense tokens, **481** pairs were attested and **510** were not.

Observations. This signal flagged many suspicious alignments at low computational cost and with far greater coverage than WSD. It also produced some false negatives when lemma/inflection or sense granularity mismatched (e.g., সম্মেলন ↔ "conference" occasionally missed), but reliably rejected clearly spurious links (e.g., আচরণ ↔ "posturing", অনিয়মিত ↔ "recrimination").

3.3 LLM-Based Verification

Finally, we prompted Gemma-3-4B-Instruct with bilingual context and the projected sense gloss, asking for a binary decision. The model *accepted* **493** and *rejected* **498** of the **991** candidates.

Observations. The LLM handled pragmatics and local context, but showed sensitivity to prompt phrasing and some run-to-run variability. It occasionally disagreed with dictionary-backed *plausible* pairs (e.g., "group" → গ্রুপের, "plan" → পরিকল্পনা) by being conservative about morphology or sense granularity. Overall it is a useful tie-breaker, best combined with symbolic signals.

4 Aggregate Picture

Core statistics	
English gold sense tokens (SE13)	**1,644**
Projected Bengali sense tokens	**991**
COMETKiwi (A1)	0.8189
COMETKiwi (A2)	**0.8488**
Filtering outcomes on 1,340 projected tokens	
Target-side WSD: agree	59
Target-side WSD: disagree	932
Dictionary: attested	481
Dictionary: not attested	510
LLM: accept	493
LLM: reject	498

Table 1: Headline results for translation comparison and filtering.

5 Discussion

What each signal contributes. WSD offers high-precision corrections when available but suffers from low BN coverage. Dictionary evidence scales well, catches many egregious links, and is the most reliable single filter in coverage vs. cost. The LLM adds flexible contextual reasoning, helping in ambiguous contexts, but introduces instability and latency.

Why A2 helps. Even a modest translation improvement reduces alignment noise (fewer literal transliterations, more natural numerals and named entities), directly lowering projection errors. COMETKiwi and manual checks confirm the expected—but small—gain.

Persistent error modes. Three issues remain salient: (i) function-word alignments leaking into content decisions; (ii) polysemy on the EN side when projection is forced to pick or skip; (iii) morphology on the BN side (e.g., case/clitics) confusing both lexicon lookups and LLM judgments.

6 Scaling and Feasibility

For much larger datasets, dictionary checks are the most scalable first-pass filter. LLM verification should be *selective* (e.g., only on pairs flagged by WSD or dictionary), with caching to reduce variance. Target-side WSD becomes more impactful as coverage improves (e.g., richer BN sense inventories). Batch MT and alignment are trivially parallelizable.

7 Conclusion

A2 translation (NLLB-200) modestly improves over A1 and yields cleaner projections. Among filters, bilingual dictionary evidence provides the best coverage–precision balance; target-language WSD offers precise but sparse confirmations; and LLM verification contributes contextual nuance with some instability. In practice, a staged filter—dictionary first, WSD when available, and LLM only on disagreements—offers a cost-effective path to higher-quality projected BN sense tags. *A simple ensemble rule* (accept if dictionary or WSD agrees; consult LLM only otherwise) can further tighten precision at a modest recall cost without heavy engineering. The overall recipe is broadly applicable to related Indic targets with minimal changes (script-aware tokenization, lemmatization, and dictionary coverage).

References

BabelfyTM. 2025. Babelfy guide. http://babelfy.org/guide. Official documentation.

Google. 2025. google/gemma-3-4b-it — hugging face. https://huggingface.co/google/gemma-3-4b-it. Model card (last updated 2025-05-30).

KakaoBrain. 2019. Github - kakaobrain/word2word: Easy-to-use word-to-word translations for 3,564 language pairs. https://github.com/kakaobrain/word2word. GitHub repository.

Meta AI. 2022. facebook/nllb-200-3.3b — hugging face. https://huggingface.co/facebook/nllb-200-3.3B. Model card.

A Manually Evaluated Examples (10)

Legend: Better = 0 (tie), 1 (A1), 2 (A2).

English	A1 (BN)	A2 (BN)	Better	Manual "Gold" (BN)
English and Teletovic were the main scorers of the 18–20 at the end of the first period.	ইংলিশ এবং টেলিটোভিচ প্রথম সময়কালের শেষে 18-20 এর প্রধান স্কোরার ছিল।	প্রথম পর্বের শেষে ইংলিশ ও তেলেটোভিচ ১৮-২০ দলের প্রধান গোলকার ছিলেন ।	0	প্রথম পর্বের শেষে ১৮–২০ স্কোরে মূল স্কোরার ছিলেন ইংলিশ ও টেলেতোভিচ ।
However, Pnini and Eidson, the best on their team, did not give up and tried to keep their team alive from the three-point line (70–77).	তবে, তাদের দলের সেরা পিনি এবং ইডসন হাল ছাড়েননি এবং তাদের দলকে তিন-পয়েন্টের লাইন (-০-7777) থেকে বাঁচিয়ে রাখার চেষ্টা করেছিলেন।	তবে তাদের দলের সেরা পনি ও আইডসন হাল ছাড়েননি এবং তিন পয়েন্টের রেখা থেকে তাদের দলকে বাঁচিয়ে রাখার চেষ্টা করেন (৭০-৭৭) ।	2	তবে, নিজেদের দলের সেরা প্পিনি ও আইডসন হাল ছাড়েননি এবং থ্রি-পয়েন্ট লাইন থেকে স্কোর করে দলকে টিকিয়ে রাখার চেষ্টা করেছেন (৭০–৭৭) ।
Wall Street closes without momentum, slowed by a strengthening dollar.	ওয়াল স্ট্রিট গতি ছাড়াই বন্ধ হয়ে যায়, একটি শক্তিশালী ডলার দ্বারা ধীর হয়ে যায়	ওয়াল স্ট্রিট বন্ধ হয়ে গেছে , ডলার বাড়ছে	1	শক্তিশালী হয়ে ওঠা ডলারের কারণে গতি হারিয়ে ওয়াল স্ট্রিট নিস্পৃহভাবে বন্ধ হয়েছে
Technology securities have thereby lost some ground.	প্রযুক্তি সিকিওরিটিগুলি এর দ্বারা কিছু স্থল হারিয়েছে।	প্রযুক্তিগত সিকিউরিটিজ এর ফলে কিছু জায়গা হারিয়ে ফেলেছে ।	0	যার ফলে প্রযুক্তি সিকিউরিটিগুলো কিছুটা অবস্থান হারিয়েছে ।
American companies walked away with stakes in just two of the 10 auctioned fields.	আমেরিকান সংস্থাগুলি 10 টি নিলাম-যুক্ত ক্ষেত্রের মধ্যে দুটিতেই অংশ নিয়ে চলে গেছে।	মার্কিন কোম্পানিগুলো নিলামের ১০টি ক্ষেত্রের মধ্যে মাত্র দুইটিতে শেয়ার নিয়ে চলে গেছে ।	2	১০টি নিলামকৃত ক্ষেত্রের মধ্যে মাত্র দুটিতে অংশীদারি নিয়ে আমেরিকান কোম্পানিগুলো ফিরেছে ।
The only one that submitted a bid lost.	একমাত্র যে একটি বিড জমা দিয়েছে।	একমাত্র ব্যক্তি যিনি একটি প্রস্তাব জমা দিয়েছিলেন তিনি হেরে গেলেন ।	0	যে একমাত্র প্রতিষ্ঠানটি দরপত্র জমা দিয়েছিল, তারা হেরে গেছে ।
Bank of America had raised some USD 19 billion from investors through convertible loans.	ব্যাংক অফ আমেরিকা রূপান্তরযোগ্য □□□□□ ণের মাধ্যমে বিনিয়োগকারীদের কাছ থেকে কিছু 19 বিলিয়ন ডলার সংগ্রহ করেছিল।	ব্যাংক অব আমেরিকা বিনিয়োগকারীদের কাছ থেকে প্রায় ১৯ বিলিয়ন ডলার ধার ধার করেছিল ।	0	ব্যাংক অব আমেরিকা কনভার্টিবল ঋণের মাধ্যমে বিনিয়োগকারীদের কাছ থেকে প্রায় ১৯ বিলিয়ন মার্কিন ডলার সংগ্রহ করেছিল ।
But that is not likely to be the last word on the issue.	তবে এটি ইস্যুতে শেষ শব্দ হওয়ার সম্ভাবনা নেই।	কিন্তু এই বিষয়ে এটাই শেষ কথা নয় ।	2	কিন্তু বিষয়টি নিয়ে এটিই শেষ কথা হওয়ার সম্ভাবনা কম ।
It wasn't a chance discovery.	এটি একটি সুযোগ আবিষ্কার ছিল না।	এটা কোন দুর্ঘটনা নয় ।	0	এটি কাকতালীয় কোনো আবিষ্কার ছিল না ।
The first is Latin America's fairly sunny mood.	প্রথমটি হ'ল লাতিন আমেরিকার মোটামুটি রৌদ্রোজ্জ্বল মেজাজ।	প্রথমটি হলো লাতিন আমেরিকার বেশ সূর্যমুখী মেজাজ ।	0	প্রথমত, লাতিন আমেরিকার বেশ আশাবাদী মনোভাব ।

Table 2: Ten manually evaluated examples. "Better": 0 = tie, 1 = A1, 2 = A2.