Projecting English Senses to Bengali: WSD, Translation, and Alignment on SE13

Mohammad Shahriar Hossain
University of Alberta
mhossai6@ualberta.ca

Ekamjot Singh
University of Alberta
ekamjot2@ualberta.ca

Abstract

This work examines cross-lingual sense projection from English to Bengali on the SemEval-2013 (SE13) subset of XL-WSD. We tag English sentences with AMuSE (Scarlini et al., 2020), translate them using `googletrans`, align words with SimAlign, and then project English BabelNet synsets onto aligned Bengali lemmas. We report the official WSD accuracy (**0.618**) and a system-level COMETKiwi quality estimate for translation (**0.8158**). Qualitative analysis highlights strengths on content words and limitations arising from multiword expressions (MWEs), named entities (NEs), and function-word alignment noise. Overall, the results illustrate when sense projection works well and where tokenization, alignment, and polysemy most commonly undermine coverage.

1 Introduction

We study cross-lingual sense projection from English to Bengali on the SemEval-2013 (SE13) subset of XL-WSD. Our workflow applies AMuSE for English word-sense disambiguation (Scarlini et al., 2020), translates the English sentences with `googletrans` (goo), aligns words using SimAlign (Jalili Sabet et al.), and then projects English BabelNet synsets onto aligned Bengali lemmas in the spirit of XL-WSD (Pasini et al., 2021). We adhere to the assignment's official formats and evaluation script for comparability.

2 WSD (English)

We tag the English side of SE13 (via XL-WSD) with AMuSE (Scarlini et al., 2020; Pasini et al., 2021). Each sentence is sent to the AMuSE HTTP API with `"lang":"EN"`. To keep a one-to-one mapping with the gold key, we lowercase token forms and compute an *nth-occurrence* index per `(sentence_id, word, POS)` so repeated tokens align deterministically to gold instance IDs. AMuSE runs with defaults; when no BabelNet ID is returned for a gold instance, we write `null` to preserve line alignment for the official scorer.

Using `evaluate_wsd.py`, overall accuracy is **0.618**.

Typical errors. (i) **POS ambiguity:** forms like *plan* are occasionally interpreted as VERB when the gold is NOUN, yielding a wrong synset. (ii) **Named entities (NEs):** proper names (e.g., *Washington*, *Copenhagen*) are often unlabeled or mapped to non-informative senses, producing `null`. (iii) **Multiword expressions (MWEs):** gold uses single-token entries (e.g., *greenhouse_gas*) while AMuSE splits them (*greenhouse* + *gas*), so the gold instance cannot be filled.

Takeaway. Performance is steady on unambiguous content words; most misses come from POS ambiguity, NE coverage, and MWE tokenization mismatches.

3 Translation (EN→BN)

We load English sentences from `se13_sentences.xlsx` (replacing `NaN` with empty strings), translate them with `googletrans.Translator` using fixed language codes (`src=en`, `dest=bn`) (goo), and save line-aligned outputs to `translations.txt`. Translation quality is estimated with COMETKiwi (`Unbabel/wmt22-cometkiwi-da`) in reference-free mode (Rei et al., 2022; Rei and IST-Unbabel), storing per-sentence scores in `translation_scores.txt` and reporting the system-level mean.

Result. The overall COMETKiwi score is **0.8158**.

Error analysis.

- **Omission of key predicate.** *"The only one that submitted a bid lost."* → MT: একমাত্র যে একটি বিড জমা দিয়েছে. The main verb "lost" is missing, changing the meaning.

- **Hallucinated token.** MT occasionally introduces a spurious string (e.g., a non-word 'Juldairs') in place of a noun phrase, suggesting unstable handling of rare contexts.
- **Mixed script and redundancy.** Phrases like "Court কোর্টহাউস" or "New নিউ" combine English and Bengali forms, producing awkward duplication.
- **Literal transliteration.** Proper names are often transliterated; in some contexts a Bengali lexical item would be more natural, affecting fluency.

Note. These issues are typical for general-purpose MT; domain-tuned EN–BN models and light post-editing (e.g., proper-noun normalization, omission checks) would help.

4 Word Alignment

We generate word alignments with SimAlign (Jalili Sabet et al.) using `bert-base-multilingual-cased` and `itermax` (defaults).

Example (`semeval2013.d000.s000`). EN: *U.N. group drafts plan to reduce emissions.* BN: মার্কিন গ্রুপ খসড়া নির্গমন হ্রাস করার পরিকল্পনা করে Alignments: `[(0,0),(1,1),(2,7),(3,6),(4,3),(5,4),(6,3)]` ⇒ (*U.N.*→মার্কিন), (*group*→গ্রুপ), (*draft*→করে), (*plan*→পরিকল্পনা), (*to*→নির্গমন), (*reduce*→হ্রাস), (*emission*→নির্গমন).

Correct: (*U.N.*, মার্কিন), (*group*, গ্রুপ), (*plan*, পরিকল্পনা), (*reduce*, হ্রাস), (*emission*, নির্গমন). *Errors:* (*draft*, করে) (morphological), (*to*, নির্গমন) (function-word noise). Note that নির্গমন is aligned twice—once correctly (with *emission*) and once incorrectly (with *to*).

Common issues: (i) function-word links; (ii) one-to-many mappings (same BN token reused); (iii) lemma–inflection mismatches. A simple post-filter that downweights stopwords and deduplicates content-word links mitigates most cases.

5 Sense Projection

We construct an intermediate file (`trans-and-ali.tsv`) containing, per sentence, the English raw text and lemmas, the Bengali translation, and SimAlign pairs (`src_idx`, `tgt_idx`). Each English token index is matched to its `instance_id` from `se13_tokens.xlsx` and then to a gold BabelNet ID in `se13.key.txt` (taking the first ID if multiple are listed). Using the alignments, we project the English BN ID to the aligned Bengali token at `tgt_idx` and write `senses.tsv` as `<bn_id><target_token>`. The final file contains **1407** rows.

Example. For `semeval2013.d003.s006`, projected senses are:

EN	BN	BN ID
two	দুটি	`bn:00021286n`
american	আমেরিকান	`bn:00014152n`
company	সংস্থা	`bn:00034265n`
reach	চুক্তিতে	`bn:00048592n`

Table 1: Projected senses for `semeval2013.d003.s006`.

Notes. Coverage is strong when alignments are present and English tokens have a single gold sense; misses mainly arise from absent/incorrect alignments, MWE or subword splits, and loss of nuance when selecting only the first BN ID for polysemous items.

6 Experimental Setup, Results, and Discussion

We use the SE13 sentence/token spreadsheets and the gold key; produced artifacts include `amuse_output.key`, `translations.txt`, `translation_scores.txt`, `alignments.txt`, and `senses.tsv`. Please review our code and scripts on GitHub.

Headline metrics	
WSD accuracy (AMuSE)	**0.618**
COMETKiwi (system-level mean)	**0.8158**
Dataset sizes	
Sentences	301
Gold instances (key lines)	1,644
Sense projection stats	
Projected senses written (rows)	**1407**

Table 2: Headline metrics, dataset sizes, and projection statistics.

Notes. Content-word senses project reliably; most residual errors trace to function-word links, tokenization mismatches, and occasional NE/MWE gaps, with general-purpose MT sometimes introducing mixed-script artifacts or omissions.

References

Amuse wsd api documentation. https://nlp.uniroma1.it/amuse-wsd/api-documentation. Accessed 2025-09-30.

googletrans: Free and unlimited google translate api for python. https://py-googletrans.readthedocs.io/en/latest/. Accessed 2025-09-30.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2855–2866, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Sree Bhattacharyya and Abhik Jana. 2022. Towards Bengali WordNet enrichment using knowledge graph completion techniques. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 75–80, Marseille, France. European Language Resources Association.

Debapratim Das Dawn. Bengali word sense disambiguation (wsd) dataset. https://www.kaggle.com/datasets/debapratimdasdawn/bengali-wsd-dataset. Kaggle dataset; Accessed 2025-10-01.

Md. Ashraful Islam, Md. Towhiduzzaman, Md. Tauhidul Islam Bhuiyan, Abdullah Al Maruf, and Jesan Ahammed Ovi. 2022. Banel: an encoder–decoder based bangla neural lemmatizer. *SN Applied Sciences*, 4:138.

Masoud Jalili Sabet, Philipp Dufter, Hinrich Schütze, and François Yvon. 2020. Simalign: High quality word alignments without parallel training data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. Simalign: High-quality word alignments without parallel training data. https://github.com/cisnlp/simalign. GitHub repository; Accessed 2025-09-30.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. https://sapienzanlp.github.io/xl-wsd/. Project website; Accessed 2025-10-01.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ricardo Rei, Luisa Coheur, Ana C Farinha, and Alon Lavie. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei and IST-Unbabel. Unbabel/wmt22-cometkiwi-da. https://huggingface.co/Unbabel/wmt22-cometkiwi-da. Hugging Face model card; Accessed 2025-09-30.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.