UAlberta at SemEval-2025 Task 2: Prompting and Ensembling for Entity-Aware Translation

Ning Shi, David Basil, Bradley Hauer, Noshin Nawal, Jai Riley
Daniela Teodorescu, John Zhang, Grzegorz Kondrak
{ning.shi,gkondrak}@ualberta.ca

Introduction

Entity-**A**ware **M**achine **T**ranslation (EA-MT) is a translation task with a focus on sentences containing **n**amed **e**ntities.

We test the following hypotheses:

Retrieval-augmented generation (RAG) & contextual learning improve LLM performance	Ensembling outputs by favoring correct NE translations improves quality
Translation quality is correlated with translation literalness	In the absence of gold named entities, a multilingual wordnet can be used as an alternative source for named entity translations.

Our methods follow two main directions:

Prompt Engineering

Using **in-context learning** and **RAG**, we develop a prompt template that focuses on translating **named entities** correctly.

Ensembling

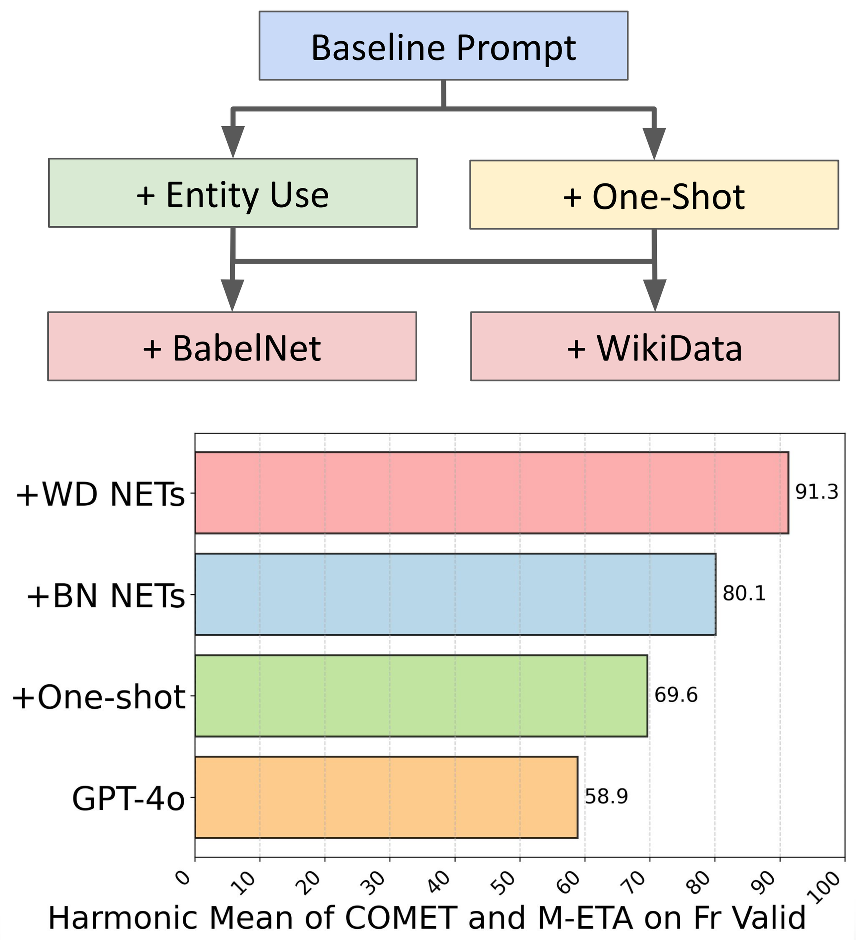
Combine inputs from **multiple MT systems**, deciding which to use based on various criteria such as **literalness** and correct translation of **named entities**.

Prompt Engineering

This method enhances the system LLM's prompt with extensions that improve depth, clarity, and detail for more accurate translations.

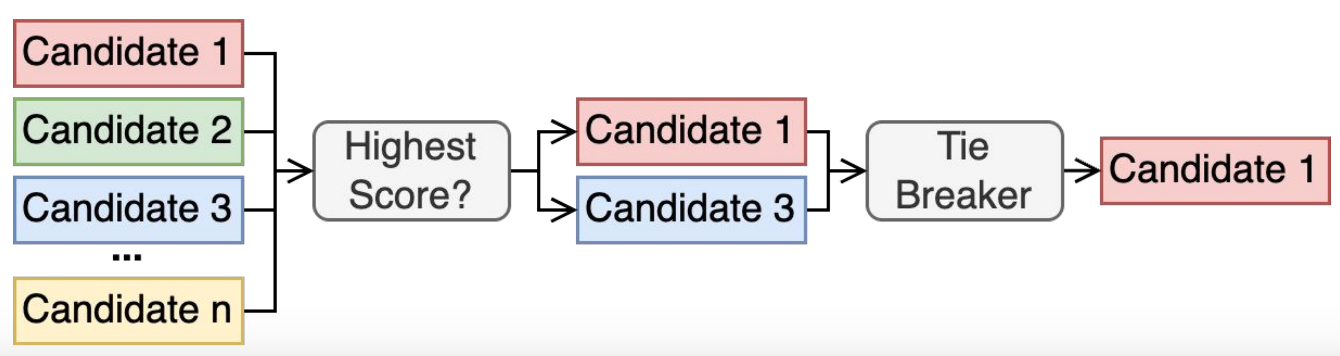
To strengthen the prompt further, other minor changes can be applied:

- Role alignment
- Anthropomorphic Prompting
- Suggestion-based Retrieval-Augmented Generation



Ensembling

- The metric for this shared task emphasizes the correct translation of entities.
- Idea: given multiple candidate translations, choose one that correctly translates the named entity in that instance.
- If multiple or no candidates correctly translate the entity, we experimented with various tie-breaking methods, including:
  - Random selection
  - Validation performance
  - Literalness, with or without semantic similarity

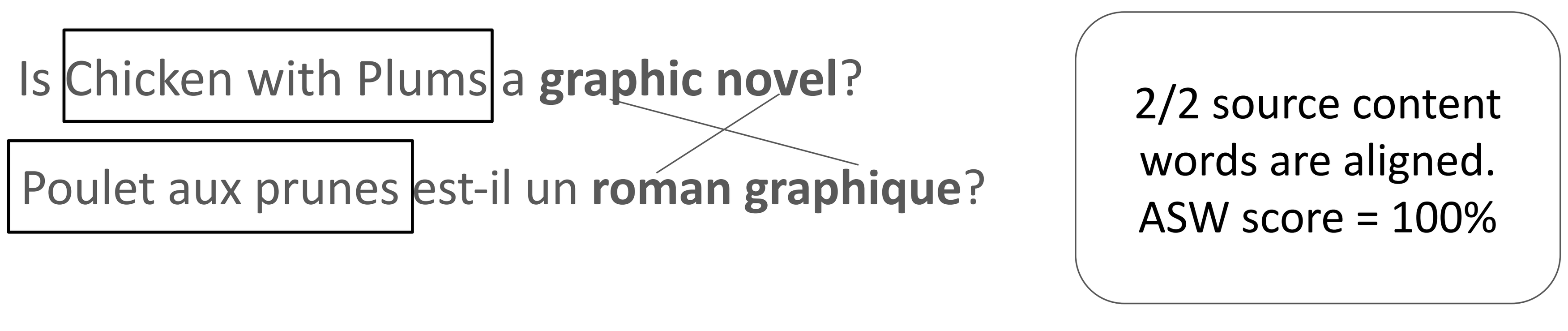


Literalness

Idea: Translation **quality** correlates with translation **literalness**.

- Use a metric of literalness to choose a translation:
- **Aligned Source Words** (ASW): The percentage of words in the source sentence aligned to words in the translated sentence.
  - Focus on content words, words with lexical meaning.
  - Ignore named entities, which are not literally translated.

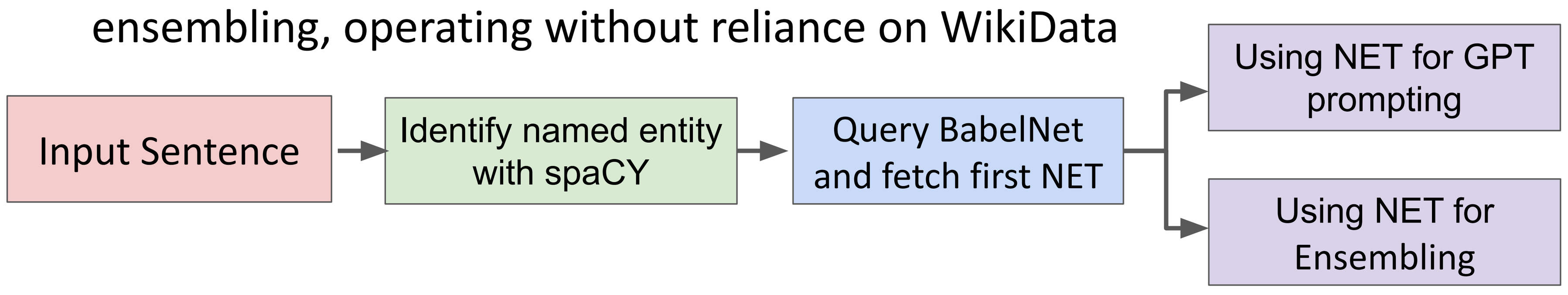
Use this score as a tiebreaking method for the named entity ensembling above.



NETs: Wikidata vs. BabelNet

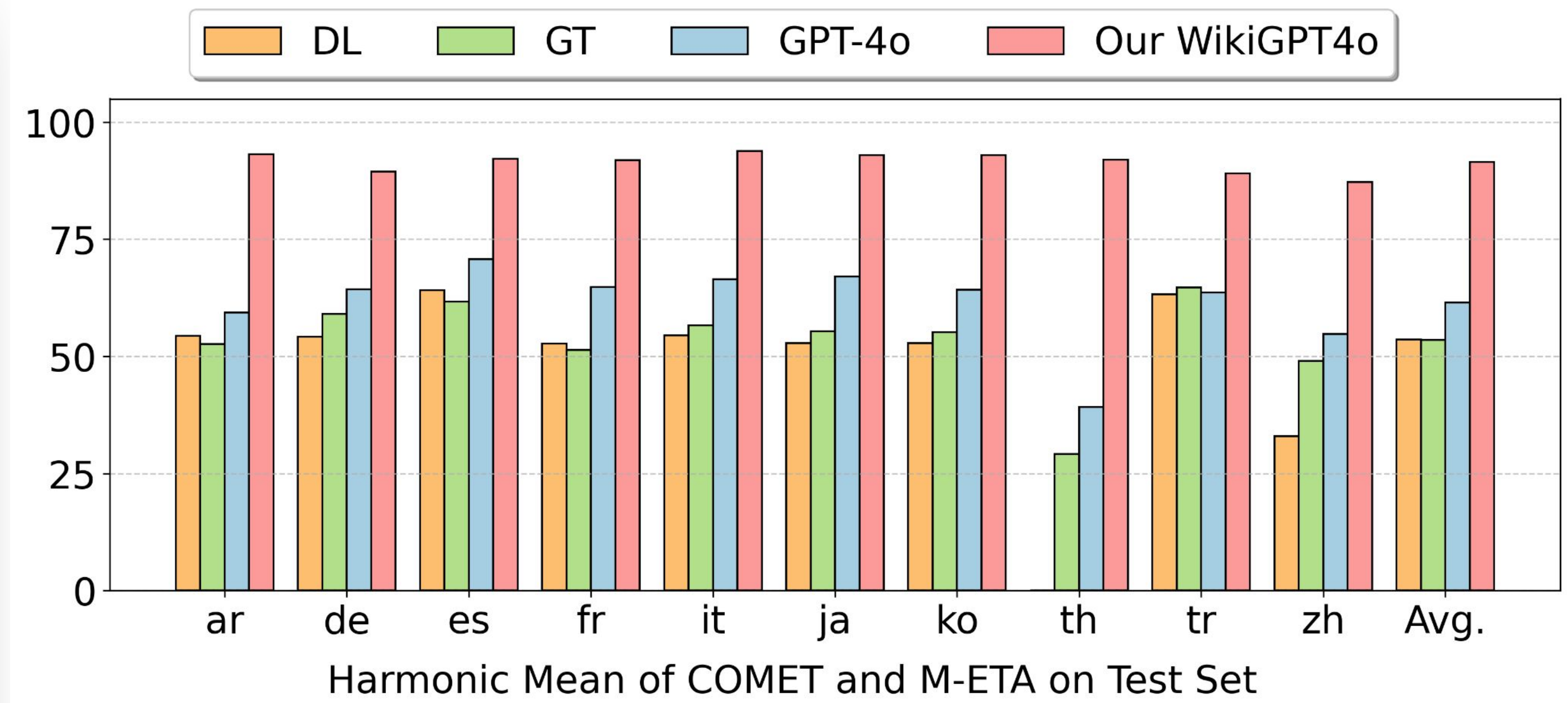
To enable our method to work without WikiData information, we retrieve Named Entity Translations (NET) from BabelNet, a multilingual semantic resource:

- Use pre-trained model to find named entities
- Query BabelNet, retrieve the first NET
- Utilising the NETs, we can now proceed with GPT prompting and ensembling, operating without reliance on WikiData



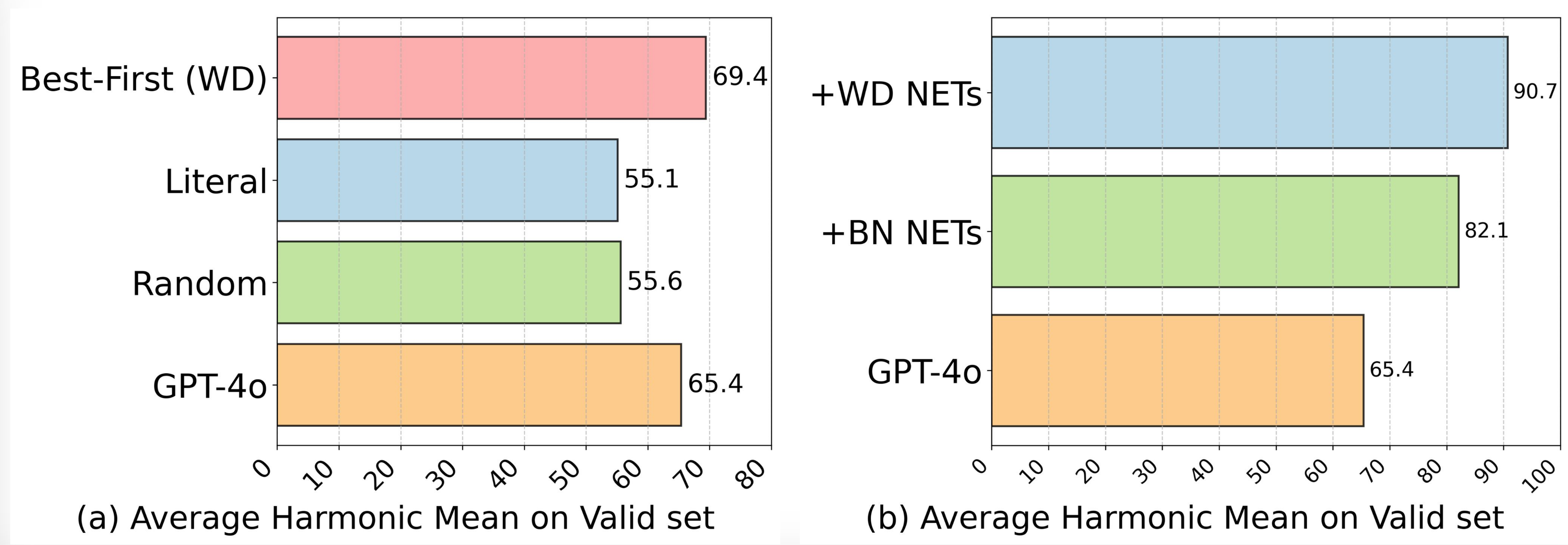
Results

The best performance is achieved by our Best-First Ensembler (WikiGPT4o) with WD Nets, which outperforms its individual component systems.



(a) While the Literal ensembler does not outperform the random baseline, our Best-First ensembler consistently achieves better results.

(b) When WikiData (WD) is unavailable, BabelNet (BN) serves as an effective alternative for supporting our NETs-based methods.



Conclusion

We validate four core hypotheses:

- Prompt engineering (RAG + in-context learning) **improves** GPT translation.
- Ensembling **boosts** performance by selecting entity-aware outputs.
- We analyze translation literalness via word alignments and find it **may not** correlate with translation quality.
- Multilingual wordnets like BabelNet are **effective alternatives** when WikiData is unavailable.

We achieved a **Top-4** out of 27 teams in overall rank, the **top** score for Arabic, and the **highest** COMET score overall.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Machine Intelligence Institute (Amii).

UNIVERSITY OF ALBERTA

amii

@SemEval 2025