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
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{ning.shi,gkondrak}@ualberta.ca

Abstract

We describe the methods used by our UAlberta team for the SemEval-2025 Task 2 on Entity-Aware Machine Translation (EA-MT). Our methods leverage large language models with prompt engineering strategies suited to this task, including retrieval augmented generation and in-context learning. Our best results overall are obtained with ensembles of multiple models, leveraging named entity knowledge in the dataset. We demonstrate that our methods work well even without gold named entity translations, by using an alternative knowledge base such as BabelNet. Finally, we provide evidence that the best translation of a given sentence is not necessarily the most literal. Our code and data are available on GitHub.

1 Introduction

This paper describes our work on SemEval 2025 Task 2: EA-MT: Entity-Aware Machine Translation (Conia et al., 2025). The EA-MT task is closely related to the well-studied task of machine translation (MT), with two key distinctions: (1) all input sentences contain a named entity, such as a location or the name of a film, and (2) the evaluation metric places a strong emphasis on the correct translation of named entities. This variant of MT is particularly challenging because named entities are often not translated literally or word-for-word (Conia et al., 2024).

The EA-MT datasets are designed to emphasize the correct translation of named entities. Each instance consists of a source English sentence, its translation in one of the target languages, a unique instance ID, the entity's Wikidata ID, the type of the entity (e.g., "Film"), and the named entity translation (NET). The data is split into four parts: training, sample, validation, and test. (As our methods are unsupervised, we do not make use of the training data.)

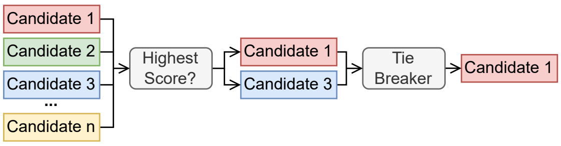

Figure 1: Our MT system ensembling template. A highest-scoring candidate translation is selected, with a tie-breaker applied if necessary.

This paper represents a further extension of our work on a comprehensive theory of lexical semantics, grounded in discrete, language-independent concepts (Hauer and Kondrak, 2020). We have applied theory-driven techniques based on multilinguality and lexical knowledge bases in our approaches to prior SemEval tasks (Hauer et al., 2020, 2021, 2022; Ogezi et al., 2023; Shi et al., 2024). The most directly relevant was the task of idiomaticity detection (Tayyar Madabushi et al., 2022), as neither idiomatic phrases nor named entities can be translated word by word (literally).

For this EA-MT task, we build upon prior work on ensembling MT systems (Figure 1). Farinhas et al. (2023) experiment with various methods for ensembling the output translations of large language models (LLMs), specifically motivated by the hallucination problem. Vernikos and Popescu-Belis (2024) leverage quality estimation systems to ensemble LLMs. However, these prior works do not specifically focus on the correct translation of named entities. Contrariwise, we incorporate NETs into both LLM prompts and ensembling systems.

In this paper, we posit and test four hypotheses: (1) Retrieval augmented generation (RAG; Lewis et al., 2020) and in-context learning (Brown et al., 2020) can improve the performance of LLMs on the EA-MT task. (2) Translation quality can be increased by ensembling the outputs of multiple MT systems via favoring NETs retrieved from a knowledge base. (3) The literalness of a sentence translation correlates with the quality of the translation. (4) In the absence of gold named entities, a multilingual wordnet can serve as a source of named entity translations.

We pursue two directions in developing our systems: prompt engineering and ensembling, which can also be used in tandem. Prompt engineering involves employing techniques for instructing LLMs such as in-context learning and RAG, which we use to develop a prompt template for LLMs with a particular focus on NETs. Our second principal direction, ensembling, consists of selecting the best translation among those produced by different MT systems given the same input. We use validation set performance, named entity identification, and literalness metrics as sources of information for our various ensembling strategies.

Our results indicate that both prompt engineering and ensembling yield improvements over baseline methods. We show that a NET-based ensembler outperforms its individual component systems. Both RAG and in-context learning improve the performance of LLMs on EA-MT. Surprisingly, using literalness to select the best translation yields no consistent improvement. In the official evaluation, our best system ranks 4th overall among 27 teams, and achieves the highest reported score for Arabic. Notably, we outperform all submitted systems on the COMET metric.

2 Methods

In this section, we describe our methods for the EA-MT task.

2.1 Prompt Engineering

To address the challenges of entity identification and phonetic or semantic adaptation, our MT system integrates RAG, in-context learning, and optimized prompt design (Liu et al., 2023) with an LLM that has strong contextual awareness and multilingual capabilities (Robinson et al., 2023).

Our GPT+NET method ("WikiGPT" in the official results table) employs a prompt template which is structured to provide additional information from external sources (Table 5 in the Appendix). This prompting strategy leverages the translations of named entities retrieved from an external knowledge base (e.g., Wikidata). The intuition is that providing explicit translation candidates guides the model toward translating the entity more accurately. In our development experiments, we found that the accuracy of our system improves when we use only the first of multiple alternative translations provided in the knowledge base.

In-context learning improves entity translations produced by LLMs by providing an example of the expected input and output. By including a high-quality reference translation, the model aligns its output with verified examples, improving fluency and correctness, and improving the handling of edge cases and linguistic nuances. In addition to integrating both retrieval and in-context learning elements, we establish a role of an "expert translator" to make the LLM adopt a professional approach to translation, and produce more precise and contextually appropriate translations.

2.2 Multi-Agent Translation

In this method, we decompose the translation process into subtasks, each handled by a different LLM agent (Guo et al., 2024). First, we tokenize the input string by identifying the longest matching substrings in BabelNet to extract potential entities. Substrings that begin with capital letters and are not located at the beginning of the sentence are passed to BabelNet to obtain all possible translations. Using RAG, these potential translations are passed to the first GPT agent, which is asked to select the best translation from the list, given the context in which it is used in the source sentence, or translate the given entity itself. The output is passed to another agent which is asked to translate the whole sentence given the already translated entities. Finally, a third agent evaluates whether the generated translation accurately conveys the source sentence; if not, it translates the source independently. We refer to this method as Multi-Agent GPT.

2.3 Named Entity Ensembling

The correct translation of named entities is given significant weight by the evaluation metric for this task. Given multiple candidate sentence translations, our primary ensembling method, which we refer to as Best-First Ensembler, assigns the maximum score to all candidates that correctly translate the named entity, and a minimal score to all other candidates. For tie-breaking purposes, we rank the MT methods according to their performance on the validation set for the language pair under consideration, and choose the candidate produced by the highest-ranked method. Our NE ensembling approach method necessarily assumes access to the correct NETs, which may be given as part of the data, or derived using available tools and resources.

2.4 Literalness

In line with our literalness hypothesis, we also test an alternative approach to ensembling which prefers the most literal translation. To obtain a literalness score, we adopt the *aligned source words* (ASW) metric of Raunak et al. (2023), which computes the proportion of words in the source sentence that are aligned to at least one word in the target sentence.[1] In our implementation, we disregard both function words and named entities. Our Literal Ensembler method selects the translation with the highest ASW.

2.5 Named Entity Translation Identification

Our prompting and named-entity ensembling methods both require at least one valid translation of the source entity. While the dataset generally includes Wikidata IDs that can be used to retrieve these NETs, we aim to maximize the generality of our methods by allowing them to operate in the absence of gold NETs. To retrieve NETs, we first apply pre-trained NER models to the source sentence, and then query a semantic knowledge base to retrieve all available translations.

3 Systems, Tools, and Resources

Our methods require multiple candidate sentence translation candidates, as well as named entity translations (NETs). For the former, we use an LLM and two commercial MT systems.

3.1 Sentence Translation

LLMs can be employed for translation by prompting them to translate an input sentence. In particular, we use the most recent version of the GPT series of LLMs, gpt-4o-2024-08-06. We set the temperature parameter to 0 to maintain translation consistency. To ensure conciseness, we limit the maximum token count per translation to 200.

DeepL is a commercial MT system. We found that DeepL may fail to translate individual sentences in a large batch. In such cases, we use CometKiwi (Rei et al., 2022) to detect the misaligned translations. DeepL does not support English-to-Thai translation.

We access Google Translate (GT) via its official API. The Chinese outputs in this task are required to be in Traditional Chinese, which is supported by both GT and DeepL. Although we specify the appropriate language code when using MT systems, we further ensure consistency in the final test set submission by converting all outputs to Traditional Chinese using the Python OpenCC library.

3.2 Named Entity Translations

Wikidata (Vrandečić and Krötzsch, 2014) is a comprehensive multilingual knowledge base, created using both manual and automated procedures, which is freely accessible via a web interface or API (Nielsen, 2020). The multilingual information in Wikidata contributes to identifying and verifying the gold standard translations for named entities. As an alternative source of NETs, we experiment with BabelNet (Navigli and Ponzetto, 2012), accessed via its Python API. Where necessary, we use spaCy to identify named entities in the text. While LLMs could be employed for NER, we found spaCy to be a more practical and sufficiently effective tool.

4 Results

We present the evaluation results of our methods across all 10 English-to-target language pairs. The principal metric for this task is the harmonic mean of COMET and M-ETA. We begin by reporting the results for our official submissions, followed by an analysis of our findings explored on the provided validation set. In addition to those already described, we report the results of three baselines: GPT as run by the shared task organizers (GPT-ST), GPT prompted without NETs (GPT-Prompt), and an ensembler which randomly selects one of the candidate translations (Random Ensembler).

4.1 Official Submissions

The results of our official submissions are reported in Table 1. Among the GPT-based methods, GPT-Prompt achieves a modest score of 61.5% when operating without external information. This shows the limitations of relying solely on its existing model capabilities without additional NET knowledge. In contrast, GPT+NET achieves an average score of 91.4%, which shows that RAG with Wikidata significantly enhances translation quality. The notable increase in the M-ETA score from 46.7% to 88.1% demonstrates the impact of leveraging external knowledge sources through RAG. Our best-performing system, Best-First Ensembler with Wikidata NETs, achieves the highest overall score of 91.5%. The slight improvement is due to

[1] For word alignment, we use SimAlign (Jalili Sabet et al., 2020) with the following hyperparameter settings: XLMR as the model, word embeddings from layer 8, bpe to define tokens, the "mai" matching method, and itermax alignment.

Method	NETs	ar	de	es	fr	it	ja	ko	th	tr	zh	Avg.
DeepL	none	54.4	54.2	64.1	52.7	54.5	52.8	52.8	n/a	63.3	33.0	53.6*
GT	none	52.6	59.1	61.7	51.4	56.6	55.4	55.2	29.2	64.7	49.0	53.5
GPT-Prompt	none	59.4	64.3	70.8	64.8	66.5	67.1	64.2	39.2	63.6	54.8	61.5
Random Ensembler	none	55.1	58.7	65.9	56.6	58.4	57.8	56.2	35.2	63.5	46.7	55.4
Best-First Ensembler	BN	61.7	63.5	69.9	62.2	63.5	67.5	63.4	38.9	67.7	57.1	61.6
Best-First Ensembler	WD	65.8	66.8	73.3	65.4	66.9	71.5	68.7	42.5	73.1	58.4	65.2
GPT+NET	WD	93.2	89.4	92.2	91.9	93.8	93.0	92.9	92.0	88.2	87.2	91.4
Best-First Ensembler	WD	93.2	89.5	92.2	91.9	93.8	93.0	93.0	92.0	89.1	87.2	91.5

Table 1: Our results on the test sets (in %), measured as the harmonic mean of COMET and M-ETA. All ensemblers use DeepL and GT, as well as the version of GPT which appears in that section of the table. The source of NETs may be Wikidata (WD), BabelNet (BN), or none. On the official leaderboard, "WikiEnsemble" corresponds to Best-First Ensembler with WD NETs, "WikiGPT4o" to GPT+NET, and "PromptGPT" to GPT-Prompt. *DeepL can not translate into Thai, so the average for DeepL is taken over only the other nine languages.

the higher M-ETA score of 88.3%, which shows the utility of ensembling that prioritizes accurate entity translation.

4.2 Prompt Engineering

To better understand the outstanding performance of our prompt template, we analyze the results of several variants of our prompt on the French validation set (Table 2). The final prompt template combines various established prompting strategies, effectively contributing to the success of our Best-First Ensembler. These experiments constitute an ablation study, allowing conclusions to be drawn from simpler variants of our key method.

Prompt templates are shown in Table 5. Starting from the official baseline prompt, we incrementally add information to the template. First, we test an increased emphasis on the entity in the prompt ("Entity Use"), and adding a single ("one shot") example of the task. Together, these prompting techniques improve the harmonic mean score of GPT's translations from 58.9% to 69.6%. Incorporating NETs from BabelNet further increases their scores to 79.1%. With the "soft NET" strategy, we allow the model to treat these entity translations as suggestions rather than constraints, which boosts performance to 80.1%. Replacing BabelNet NETs with Wikidata NETs produces the best score by far, 91.3%. These results prove that, unlike conventional translation services, LLMs-based translations can be refined by prompt design, demonstrating the benefits of techniques like RAG and in-context learning.

4.3 Additional Findings

In this section, we discuss the multi-agent translation method (Section 2.2), and describe some peculiarities pertaining to Wikidata.

Prompt	NETs	M-ETA	COMET	HM
GPT-ST	none	44.1	88.9	58.9
+Entity Use	none	56.1	91.6	69.6
+One-shot*	none	56.1	91.6	69.6
+Entity Use	BN	69.1	92.6	79.1
+One-shot	BN	69.2	92.4	79.1
+Soft NETs	BN	70.4	92.8	80.1
+Soft NETs*	WD	88.5	94.2	91.3

Table 2: Results of different prompts on the French (fr) validation set. Methods marked with * correspond to the version used in our final submission (GPT-Prompt and GPT+NET).

We did not submit the Multi-Agent GPT results, as our implementation of this method underperforms the one-shot prompt. This is principally due to the difficulty in prompting LLMs to consistently follow instructions for intermediate steps. The issues that reduce the reliability and effectiveness of the multi-agent approach include: (1) copying the English NE instead of translating it, (2) answering the source question instead of translating it, (3) error propagation between the agents, and (4) the presence of lowercase entities in the dataset.

When analyzing our translations, we observe that the provided gold reference translations sometimes conflict with the information we retrieved from Wikidata. We found that these discrepancies are due to two factors. First, the Wikidata IDs provided in the dataset do not always match those we retrieved from the Wikidata API or web interface. For example, the Arabic entity "Andalusian Mosque" is linked to ID "Q12204195" in the task dataset, whereas its correct entry in Wikidata is "Q3324925". Second, the entity translations in the dataset have been refined manually (Conia et al.,

Method	NETs	ar	de	es	fr	it	ja	ko	th	tr	zh	Avg.
DeepL	none	52.4	54.1	64.2	56.3	60.9	47.4	55.1	n/a	58.0	32.6	53.4*
GT	none	47.8	58.5	63.0	59.2	60.2	52.1	59.0	26.6	60.8	50.4	53.8
GPT-ST	none	53.6	57.3	66.9	58.9	61.1	60.0	62.8	27.5	53.5	50.3	55.2
GPT-Prompt	none	59.1	68.3	76.4	69.6	72.0	69.8	72.9	41.0	65.4	59.0	65.4
Random Ensembler	none	53.0	60.7	67.8	62.0	64.5	55.3	62.0	33.4	60.2	37.6	55.6
Literal Ensembler	none	50.9	59.9	67.8	61.6	64.5	55.7	60.3	31.2	59.2	39.4	55.1
Best-First Ensembler	WD	65.6	72.0	78.3	72.8	75.3	73.7	77.0	44.8	72.5	61.9	69.4
Multi-Agent GPT	BN	71.7	63.7	70.7	63.6	70.3	72.2	79.6	56.8	70.3	68.7	68.7
GPT+NET	BN	87.5	77.8	79.5	80.1	84.0	85.4	86.0	77.3	82.3	81.2	82.1
GPT+NET	WD	93.6	89.6	92.6	91.3	94.7	93.1	91.7	91.4	86.1	83.4	90.7
Best-First Ensembler	WD	93.8	89.9	92.6	91.3	94.7	93.2	91.7	91.6	86.2	83.4	90.8

Table 3: Our results on the validation sets (in %), measured as the harmonic mean of COMET and M-ETA. All ensemblers use DeepL and GT, as well as the version of GPT which appears in that section of the table. For the validation sets, no conversion from Simplified to Traditional Chinese was applied.

2024), resulting in discrepancies between the gold translations and Wikidata information. For example, in Wikidata the entity "Q1761410" is translated as "Dark Night of the Soul" in both English and German, with the German entry simply copying the English name. These cases show that some errors in our translations actually arise from dependence on Wikidata. We found that 88.6% of the translations in the validation set and 88.1% of the translations in the test set were consistent with the gold translations. More details can be found in Table 4 in the Appendix.

4.4 Hypotheses and Evidence

In Section 1, we put forward four hypotheses related to the EA-MT task. In this section, we assess, for each hypothesis, what conclusions we can draw about these hypotheses based on our empirical findings. In some cases, we have carried out additional experiments after the official submission period to obtain additional data and resolve open questions. The results in Table 3 in the Appendix present our extended evaluation results on the validation set.

First, our experiments confirm that RAG and in-context learning significantly improve LLMs for the EA-MT task. Without NETs, applying prompt engineering with in-context learning improves the performance of GPT by an average of 10% over the reproduced official GPT baseline. With NETs, using either BabelNet or Wikidata further boosts performance to 82.1% and 90.7%, respectively, demonstrating the effectiveness of integrating external knowledge sources in enhancing entity translation accuracy.

Second, our results confirm that ensembling multiple translation outputs with named entity retrieval consistently improves translation quality. Using DeepL, GT, and GPT-Prompt as base translators, Best-First Ensembler increases the performance of best single translator GPT-Prompt from 65.4% to 69.4%. Even though GPT+NET already achieves a high score with access to Wikidata, our Best-First Ensembler further enhances its performance, reaching 90.8%, the highest among all our systems. This reinforces the effectiveness of prioritizing candidates that contain correct NETs.

Third, our experiments with literalness-based ensembling provide no evidence that favoring literal translations improves ensemble quality overall. In fact, we observe a drop in performance with respect to the random baseline. To further explore this surprising finding, we analyzed the literalness scores of the provided gold sentence translations. We found that many gold translations are in fact less literal than our candidate translations. On average, the ASW score of the first gold translation for each sample in the validation dataset is 82.2%, whereas the mean ASW for our candidate translations is 83.0% for GPT, 83.5% for DeepL, and 84% for GT. For example, GPT translation of *"blood sugar levels"* as *"les niveaux de sucre dans le sang"* appears more literal than the gold translation *"glycémie"*. Taken together, these experiments do not provide evidence for our third hypothesis linking literalness to translation quality; the most literal translation, we find, is not necessarily the best.

Finally, our results confirm that, in the absence of gold named entity links, using BabelNet for NETs instead of Wikidata still consistently improves performance, raising overall scores from the 60% range to approximately 80%. For the best single translator, GPT-Prompt, integrating BabelNet significantly enhances its performance, increasing the average score from 65.4% to 82.1%. These results underline the importance of explicit entity translation and demonstrate that multilingual wordnets can serve as effective alternatives when knowledge resources like Wikidata are unavailable.

These findings validate three of the four hypotheses proposed at the outset, reinforcing the importance of prompt engineering, ensembling strategies, and external knowledge integration in improving modern entity-aware machine translation.

5 Conclusion

After extensive experimentation on the EA-MT datasets, we conclude that three of four hypotheses formulated in Section 1 are well-supported by our empirical results. Our work provides new evidence supporting the use of external knowledge bases in semantic tasks, and prompts further exploration of word-level analysis, especially as it corresponds to literalness. Finally, we note that our highest-scoring methods were ranked at or near the top of the official leaderboards for several languages and categories.

Limitations

Our system integrates several components, ensembling multiple MT systems and retrieving NETs from external knowledge bases. This design not only contributes to strong performance but also increases computational requirements. In fixed dataset settings, the process remains manageable through preprocessing and sequential execution.

Scalability is an important factor when applying our system to larger datasets or new domains. While the current setup works well for the shared task, broader use may benefit from optimizations. Leveraging multiple translation systems and external lookups can lead to increased latency and higher resource consumption, especially in high-throughput or multilingual scenarios.

Real-time or latency-sensitive applications may require further adjustments. Model ensembling and external database queries are less practical for instant translation. In addition, depending on structured resources like Wikidata or BabelNet may reduce the system effectiveness in domains or languages with limited coverage.

Acknowledgments

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

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Bradley Hauer, Hongchang Bao, Arnob Mallik, and Grzegorz Kondrak. 2021. UAlberta at SemEval-2021 task 2: Determining sense synonymy via translations. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 763–770, Online.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 263–269, Barcelona (online).

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, and Grzegorz Kondrak. 2022. UAlberta at SemEval 2022 task 2: Leveraging glosses and translations for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 145–150, Seattle, United States.

Bradley Hauer and Grzegorz Kondrak. 2020. Synonymy= translational equivalence. *arXiv preprint arXiv:2004.13886*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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.

Finn Nielsen. 2020. Lexemes in Wikidata: 2020 status. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 82–86, Marseille, France. European Language Resources Association.

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, and Grzegorz Kondrak. 2023. UAlberta at SemEval-2023 task 1: Context augmentation and translation for multilingual visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2043–2051, Toronto, Canada.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Ning Shi, Senyu Li, Guoqing Luo, Amirreza Mirzaei, Ali Rafiei, Jai Riley, Hadi Sheikhi, Mahvash Siavashpour, Mohammad Tavakoli, Bradley Hauer, and Grzegorz Kondrak. 2024. UAlberta at SemEval-2024 task 1: A potpourri of methods for quantifying multilingual semantic textual relatedness and similarity. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1798–1805, Mexico City, Mexico.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Giorgos Vernikos and Andrei Popescu-Belis. 2024. Don't rank, combine! Combining machine translation hypotheses using quality estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12087–12105, Bangkok, Thailand. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

A Appendix

Lang.	**Total Instances**	**Gold-WD Label Match**	**WD-Label Match%**	**WD Other Trans. Match%**	**Overall Match**	**Overall Match%**	**Mismatch%**
ar	722	666	92.2	10.9	680	94.2	05.8
de	731	647	88.5	20.5	688	94.1	05.9
es	739	673	91.1	31.0	693	93.8	06.2
fr	724	668	92.3	26.5	700	96.7	03.3
it	730	695	95.2	13.4	701	96.0	04.0
ja	723	669	92.5	14.5	684	94.6	05.4
ko	745	660	88.6	08.9	675	90.6	09.4
th	710	657	92.5	12.8	665	93.7	06.3
tr	732	626	85.5	08.2	644	88.0	12.0
zh	722	484	67.0	11.8	538	74.5	25.5
Summary	7278	6445	88.6	15.9	6668	91.6	08.4

Lang.	**Total Instances**	**Gold-WD Label Match**	**WD-Label Match%**	**WD Other Trans. Match%**	**Overall Match**	**Overall Match%**	**Mismatch%**
ar	4547	4225	92.9	10.3	4282	94.2	05.8
de	5876	5216	88.8	16.9	5452	92.8	07.2
es	5338	4864	91.1	28.6	5027	94.2	05.8
fr	5465	5080	93.0	26.7	5234	95.8	04.2
it	5098	4833	94.8	15.7	4933	96.8	03.2
ja	5108	4550	89.1	13.0	4722	92.4	07.6
ko	5082	4534	89.2	08.0	4599	90.5	09.5
th	3447	3168	91.9	10.3	3237	93.9	06.1
tr	4473	3906	87.3	12.8	4014	89.7	10.3
zh	5182	3316	64.0	12.0	3741	72.2	27.8
Summary	49616	43692	88.1	15.8	45241	91.2	08.8

Table 4: The match and mismatch percentages between Gold and Wikidata translations for the validation and test set, in that order. The columns represent: (1) Lang. - language codes, (2) Total Instances - overall number of NEs, (3) Gold-WD Label Match - number of gold translations that agree with what we retrieve from Wikidata (main translations only), (4) WD-Label Match% - ratio of Gold-WD Label Match to Total Instances (5) Other Trans. Match% - the same ratio, when WikiData's alias translations are used instead, (6) Overall Match - same as Gold-WD Label Match, but with alias translations included, (7) Overall Match% - ratio of Overall Match to Total Instances, (8) Mismatch% - 1 minus Overall Match% (percent of instances where no WikiData translation matches the gold translation of the NE).

Prompt	**NETs**	
GPT-ST	none	You are an expert translator. Translate from ***source_language*** to ***target_language***. Provide only the translation without explanations.
+ Entity Use	none	You are an expert translator. Translate from ***source_language*** to ***target_language*** while preserving meaning and proper entity translation. Identify the named entity in the ***source_language*** sentence and search for its translations in ***target_language*** from Wikidata, and use the named entity translation in the translated sentence. Provide only the translated text without explanations.
+ One-shot* (GPT-Prompt)	none	You are an expert translator. Translate from ***source_language*** to ***target_language*** while preserving meaning and proper entity translation. Refer to the example translation for consistency: Source: ***source_sentence*** Target: ***target_sentence*** Provide only the translated text without explanations.
+ Entity Use	BN	You are an expert translator. Translate from ***source_language*** to ***target_language*** while preserving meaning and proper entity translation. Identify the named entity in the ***source_language*** sentence and translate it accurately as ***ne_translation*** in ***target_language*** Then, provide a full translation of the sentence into ***target_language***, ensuring the named entity is translated exactly as specified. Provide only the translated text without explanations.
+ One-shot	BN	You are an expert translator. Translate from ***source_language*** to ***target_language*** while preserving meaning and proper entity translation. Identify the named entity in the ***source_language*** sentence and translate it accurately as ***ne_translation*** in ***target_language*** Then, provide a full translation of the sentence into ***target_language***, ensuring the named entity is translated exactly as specified. Refer to the example translation for consistency: Source: ***source_sentence*** Target: ***target_sentence*** Provide only the translated text without explanations.
+ Soft NETs	BN	You are an expert translator. Translate from ***source_language*** to ***target_language*** while preserving meaning and proper entity translation. A possible translation for the entity in the sentence is ***ne_translation***. Use this if you think it is correct. Refer to the example translation for consistency: Source: ***source_sentence*** Target: ***target_sentence*** Provide only the translated text without explanations.
+ Soft NETs* (GPT+NET)	WD	You are an expert translator. Translate from ***source_language*** to ***target_language*** while preserving meaning and proper entity translation. The named entity ***named_entity*** should be translated appropriately, considering the best contextual translation. Use the most suitable translation from: ***all_translations***, with the first one being the most likely. Refer to the example translation for consistency: Source: ***source_sentence*** Target: ***target_sentence*** Provide only the translated text without explanations.

Table 5: Versions of prompts used in GPT translation. Variables are provided in ***italic bold*** font.