Exploiting Word Sense Disambiguation in Large Language Models for Machine Translation

Van-Hien Tran, Raj Dabre, Hour Kaing, Haiyue Song, Hideki Tanaka, Masao Utiyama
National Institute of Information and Communications Technology (NICT)
{tran.vanhien, raj.dabre, hour_kaing}@nict.go.jp
{haiyue.song, hideki.tanaka, mutiyama}@nict.go.jp

Abstract

Machine Translation (MT) has made great strides with the use of Large Language Models (LLMs) and advanced prompting techniques. However, translating sentences with ambiguous words remains challenging, especially when LLMs have limited proficiency in the source language. This paper introduces two methods to enhance MT performance by leveraging the word sense disambiguation capabilities of LLMs. The first method integrates all the available senses of an ambiguous word into the prompting template. The second method uses a pre-trained source language model to predict the correct sense of the ambiguous word, which is then incorporated into the prompting template. Additionally, we propose two prompting template styles for providing word sense information to LLMs. Experiments on the HOLLY dataset demonstrate the effectiveness of our approach in improving MT performance.

1 Introduction

Semantic ambiguity has long posed a significant challenge in MT. Despite rapid advancements in Neural Machine Translation (NMT), effectively disambiguating and translating ambiguous words remains an unresolved issue. The advent of decoder-only large language models (LLMs) such as the GPT series (Achiam et al., 2023), LLaMA (Touvron et al., 2023a,b), and Gemma (Mesnard et al., 2024) has shown exceptional capabilities in various natural language processing tasks, including MT. These LLMs have emerged as promising alternatives, offering performance comparable to traditional NMT models and introducing new paradigms for controlling target outputs.

However, due to their predominant pre-training on English-centric language datasets (Naveed et al., 2023), LLMs may lack proficiency in low-resource languages (Tran et al., 2023), making it challenging for them to accurately translate source sentences containing ambiguous words in these languages (Campolungo et al., 2022; Nambi et al., 2023). This issue is particularly pronounced in small and moderate-sized models (2B, 7B, or 13B) (Scao et al., 2022; Lu et al., 2024; Vo, 2024). In this study, we investigate the translation capabilities of such LLMs in handling ambiguous words through prompting techniques, without relying on additional training data. In addition, we present two methods to take advantage of the word-sense disambiguation (WSD) abilities of LLMs, thus enhancing MT performance.

The first method integrates all possible senses of the ambiguous word from a dictionary into the prompting template, encouraging LLMs to use their internal WSD capabilities to select the appropriate word sense, thus improving translation quality. The second method utilizes an external decoder-only language model pre-trained on a large set of source language data. This model evaluates the perplexities of all sense definitions from a dictionary in the source language and predicts the correct sense with the lowest perplexity. The predicted sense is then incorporated into the prompting template to aid the LLMs in the translation process. Besides, we propose two prompting template styles for each method: *Natural Language Style* and *Tagging Style*.

Our contributions are as follows:

(a) We introduce two methods that leverage the WSD capabilities of LLMs to enhance MT performance on sentences with ambiguous words.

(b) We present two prompting template styles for each method, integrating word sense information into LLMs to address MT task.

(c) Experiments on the HOLLY dataset (Baek et al., 2023) demonstrate the effectiveness of our approach in utilizing WSD capabilities of LLMs, leading to improved MT performance.

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 135–144
January 20, 2025.

2 Related Work

Zero-shot and few-shot prompting have become essential techniques for leveraging LLMs in MT. Zero-shot prompting asks the model to translate directly without examples, while few-shot prompting provides a few examples to guide the model through in-context learning (Brown et al., 2020). Previous works (Radford et al., 2019; Jiao et al., 2023) have shown that both methods can achieve competitive results without extensive fine-tuning. Although fine-tuning LLMs in specific language pairs can improve MT (Zhang et al., 2023), it demands computational resources and annotated data.

More related to our work, Pilault et al. (2023) proposed interactive-chain prompting, a prompt-based interactive multi-step computation technique that first resolves cross-lingual ambiguities in the input queries and then performs conditional text generation. Iyer et al. (2023) presented two techniques to improve the disambiguation abilities of LLMs, including in-context learning and fine-tuning. The former involves providing similar ambiguous contexts in the prompt, while the latter involves fine-tuning LLMs on carefully curated ambiguous datasets through low-rank adaptation. Unlike these approaches, our approach takes advantage of the WSD capabilities of LLMs to improve MT without additional fine-tuning.

3 Our Method

Given a source sentence containing the ambiguous word in language X, our goal is to use LLMs to accurately translate the sentence into language Y. Figure 1 illustrates our approach using the pair (X,Y) as (Korean, English). Following Xu et al. (2024), we use a basic prompting format: "Translate this from Korean to English:\nKorean:<source sentence>\nEnglish:" on LLMs, as illustrated in Block 1 of Figure 1.

To enhance LLMs' ability to translate sentences containing ambiguous words, we use a dictionary to gather all possible senses of the ambiguous word. For example, in Block 2 of Figure 1, the word '연기' has three distinct senses, each with an English translation and a definition in Korean. We present two methods to exploit this information for LLMs.

All Senses-based Prompting. This method incorporates all potential senses of the ambiguous word into the prompting template, utilizing two distinct styles: *Natural Language Style* (*NLS*) and *Tagging Style* (*TS*). By providing such information, it exploits the WSD ability of LLMs for ambiguous words, thereby improving MT accuracy.

Figure 1: The overall framework.

As shown in Block 3 of Figure 1, for the *NLS*, we provide all senses of the word '연기' in a natural language format: "Hint: '연기' means 'smoke' or 'delay' or 'acting'." In contrast, the *TS* uses tags to convey the word sense information. For instance, the ambiguous word '연기' is followed by the tag "<w>smoke, delay, acting</w>".

One Predicted Sense-based Prompting. This method predicts the most relevant sense of an ambiguous word in a source sentence and provides this prediction to LLMs, instead of listing all possible senses. We use a decoder-only language model pre-trained exclusively in the source language. For example, let $\mathcal{M}$ be a decoder-only model trained solely in Korean. Due to its lack of proficiency in the target language, the model $\mathcal{M}$ is unable to directly translate the input sentence from the source language to the target language.

Given $\mathcal{M}$'s deep understanding of Korean, we leverage it to predict the correct sense of the ambiguous word. We use the template $\mathcal{T}$: "문맥 'A' 에서 키워드 'B'는 다음을 의미합니다. " (translated as: "In the 'A' context, 'B' means: "), where A is the source sentence and B is the ambiguous word. Assuming that B has K distinct senses from a Korean-English dictionary, our objective is to predict the correct sense of B in A.

For each candidate sense S_j, we combine $\mathcal{T}$ with its Korean definition to create a full statement. This statement is then tokenized into N tokens: $w_1, w_2, \ldots, w_{N_1}, w_{N_1+1}, \ldots, w_N$. The first N_1 tokens come from $\mathcal{T}$, while the rest are from the sense definition. We calculate the perplexity for each candidate using two various methods. The first method calculates perplexity over all N tokens:

$$\mathrm{PPL}_{\text{full}} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log P_{\mathcal{M}}(w_i \mid w_1, \ldots, w_{i-1})\right)$$

Meanwhile, the second method calculates perplexity only over the $(N - N_1)$ tokens of the sense definition in the full statement:

$$\mathrm{PPL}_{\text{def}} = \exp\left(-\frac{1}{N-N_1}\sum_{i=N_1+1}^{N} \log P_{\mathcal{M}}(w_i \mid w_1, \ldots, w_{i-1})\right)$$

Here, $P_{\mathcal{M}}(w_i \mid w_1, \ldots, w_{i-1})$ is the probability of token w_i given its preceding context as estimated by the model $\mathcal{M}$. After obtaining the perplexity scores for all K candidate senses of the ambiguous word, the sense with the lowest perplexity is selected as the most likely correct sense: $\hat{S} = \arg\min_{j\in\{1,\ldots,K\}} \mathrm{PPL}(S_j)$.

We incorporate the above predicted sense into the prompting template, as shown in Block 4 of Figure 1, using two styles: *NLS* and *TS*, similar to "All Senses-based Prompting". By providing a single, highly reliable predicted sense, we aim to help LLMs better understand ambiguous words.

4 Experiments

4.1 Dataset and Settings

Dataset. We evaluate our approach using the HOLLY benchmark test set (Baek et al., 2023). It includes 600 high-quality Korean-to-English translation test examples, where each source sentence contains one homograph word. Homographs are words that have the same form but multiple different senses, which can lead to ambiguity without context. However, the specific context of each source sentence typically clarifies the correct sense. Out of the 600 examples, 300 are positive test examples in which the correct sense of the homograph is labeled. Refer to Appendix A for details.

Settings. We evaluate our approach on five LLMs using 1-shot and 3-shot learning. The models include Gemma-2B[1], Gemma-7B[2], LlaMA-2-7B[3], LlaMA-2-13B[4], and LlaMA-3-8B[5], all available on Huggingface[6]. We keep all LLM parameters frozen during the experiments.

For text generation, we use non-sampling greedy decoding, a maximum of 100 new tokens, and BF16 precision. Each experiment runs on a machine with eight NVIDIA Tesla V100 Volta 32GB GPUs and a maximum runtime of 6 hours. The chrF++ metric[7] (Popović, 2017) is used to evaluate MT. We utilize the available pre-trained Korean language model Polyglot-Ko-12.8B[8] as $\mathcal{M}$ introduced in Section 3. In scenarios where such pre-trained source-side models are unavailable, we propose pre-training these models using accessible monolingual datasets.

We also refer to the Korean-English dictionary from the National Institute of Korean Language[9]. Besides, we prepare three fixed examples to use for prompting with 1-shot and 3-shot learning. They are provided in Table 4.

4.2 Results and Analysis

Accuracy of the Sense Prediction Module. Our method, "One Predicted Sense-based Prompting", features a sense prediction module that identifies the most relevant sense of an ambiguous word based on its context. We evaluate the accuracy of this module on 300 positive examples of the HOLLY test set. Table 1 shows that both $\mathrm{PPL}_{\text{full}}$ and $\mathrm{PPL}_{\text{def}}$ obtain high accuracy, with $\mathrm{PPL}_{\text{def}}$ reaching 91.67 percent. As each ambiguous word in the test examples has at least two different senses, these results highlight the pre-trained model's strong proficiency in Korean and its effectiveness in reliably predicting word senses in context.

[1]https://huggingface.co/google/gemma-2b
[2]https://huggingface.co/google/gemma-7b
[3]https://huggingface.co/meta-llama/Llama-2-7b-hf
[4]https://huggingface.co/meta-llama/Llama-2-13b-hf
[5]https://huggingface.co/meta-llama/Meta-Llama-3-8B
[6]https://huggingface.co/
[7]nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.4.1
[8]https://huggingface.co/EleutherAI/polyglot-ko-12.8b
[9]https://krdict.korean.go.kr

Ours	Accuracy
$\mathrm{PPL}_{\text{full}}$	87.78
$\mathrm{PPL}_{\text{def}}$	91.67

Table 1: Accuracy of the sense prediction module

	Model	Baseline	All Senses		Predicted Sense	
			NLS	*TS*	*NLS*	*TS*
1-shot	Gemma-2B	31.73	34.60	30.72	**34.79**	32.55
1-shot	Gemma-7B	33.22	35.55	35.67	36.43	**37.26**
1-shot	LlaMA-2-7B	22.63	28.82	29.16	**30.42**	30.36
1-shot	LlaMA-2-13B	42.51	45.09	44.71	45.60	**46.11**
1-shot	LlaMA-3-8B	44.05	46.85	45.83	47.11	**47.40**
3-shot	Gemma-2B	30.33	31.47	28.94	**32.62**	30.55
3-shot	Gemma-7B	35.49	37.12	37.17	37.63	**38.29**
3-shot	LlaMA-2-7B	24.86	30.29	30.81	**31.54**	31.06
3-shot	LlaMA-2-13B	43.40	44.91	45.05	45.69	**46.38**
3-shot	LlaMA-3-8B	44.35	46.94	45.76	**47.22**	47.15

Table 2: Performance on MT of the different prompting methods using ChrF++. *NLS* and *TS* stand for *Natural Language Style* and *Tagging Style*, respectively.

Performance on MT. With the high accuracy of the sense prediction module, we evaluate performance on MT of our "One Predicted Sense-based Prompting" method against other approaches, using the entire HOLLY test set. Table 2 presents the results, where **Baseline**, **All Senses**, and **Predicted Sense** correspond to "Basic Prompting", "All Senses-based Prompting", and "One Predicted Sense-based Prompting", respectively. Four key findings from Table 2 are highlighted below.

First, the **Baseline** results indicate that performance generally improves in the 3-shot scenario compared to the 1-shot scenario for all models, except for the Gemma-2B model, which shows a slight decrease of 1.4 points. This trend highlights the effectiveness of few-shot learning, as providing more examples typically enhances model performance, though the degree of improvement varies across different models. Notably, LlaMA-2-7B has the lowest performance in both scenarios, while LlaMA-3-8B achieves the highest performance among the five models.

Second, the best performance of **All Senses** and **Predicted Sense** across all five models in both 1-shot and 3-shot scenarios shows a significant improvement over the **Baseline**. This consistent enhancement suggests that providing word sense information for ambiguous words in source sentences greatly aids in generating accurate translations. Notably, our approach yields the most substantial improvement with LlaMA-2-7B in both 1-shot and 3-shot scenarios, even though this model has the lowest **Baseline** performance. For instance, in the 1-shot scenario with LlaMA-2-7B, **All Senses** and **Predicted Sense** improve the **Baseline** by 6.53 points and 7.79 points, respectively. This indicates that word sense information is particularly crucial for LLMs with limited source language abilities, as it significantly enhances their translation accuracy.

	Model	Baseline	Predicted Sense		Gold Sense	
			NLS	*TS*	*NLS*	*TS*
1-shot	Gemma-2B	33.13	35.12	33.16	35.40	33.63
1-shot	Gemma-7B	35.15	37.28	37.53	37.61	37.86
1-shot	LlaMA-2-7B	23.21	31.05	30.81	31.67	31.60
1-shot	LlaMA-2-13B	43.33	45.95	46.63	46.15	46.95
1-shot	LlaMA-3-8B	45.06	47.14	47.62	47.58	48.01
3-shot	Gemma-2B	32.26	33.75	31.10	33.83	31.33
3-shot	Gemma-7B	37.40	38.59	39.68	38.83	40.09
3-shot	LlaMA-2-7B	25.72	32.42	32.01	32.93	32.35
3-shot	LlaMA-2-13B	44.04	45.91	46.28	46.13	46.81
3-shot	LlaMA-3-8B	45.40	47.41	47.18	47.91	47.70

Table 3: Impact of the Sense Prediction Accuracy on MT using ChrF++ over 300 samples. *NLS* and *TS* stand for *Natural Language Style* and *Tagging Style*, respectively.

Third, in both 1-shot and 3-shot scenarios, **Predicted Sense** consistently outperforms **All Senses** across all five models on both *NLS* and *TS*. On average, it improves the ChrF++ scores by 0.74 points on *NLS* and 1.33 points on *TS*. The most significant improvements are observed with Gemma-2B on *TS*, where **Predicted Sense** surpasses **All Senses** by 1.83 points in the 1-shot scenario and 1.62 points in the 3-shot scenario. These results highlight the advantage of exploiting the WSD capability of an external pre-trained source language model to provide the relevant sense of ambiguous words in context, thereby enhancing the performance of general-purpose LLMs in MT.

Last, we compare the performance differences between *NLS* and *TS* for both **All Senses** and **Predicted Sense**. For the small-sized LLM, Gemma-2B, *NLS* proves more effective than *TS* in both 1-shot and 3-shot scenarios, likely because Gemma-2B better understands and uses word sense information in natural language form. Conversely, for the moderate-sized LLMs (the four remaining models), the differences between *NLS* and *TS* are not significant in either 1-shot or 3-shot scenarios. These models effectively understand word sense information regardless of the format, achieving competitive MT performance with both *NLS* and *TS*.

Impact of the Sense Prediction Accuracy on MT. We examine how the accuracy of the sense prediction in our "One Predicted Sense-based Prompting" method affects MT performance using 300 positive test examples from the HOLLY test set. Table 3 shows the results, comparing **Baseline** (Basic Prompting), **Predicted Sense** (One Predicted Sense-based Prompting), and **Gold Sense** (One Gold Sense-based Prompting).

We contrast MT performance between **Predicted Sense** with 91.67% accuracy (from Table 1) and **Gold Sense** with 100% accuracy. The results in Table 3 demonstrate consistent improvements when using **Gold Sense** compared to **Predicted Sense** across both *NLS* and *TS* settings. For every model and scenario, **Gold Sense** yields higher scores than **Predicted Sense**, even if the improvements are sometimes small. This shows that providing more accurate word sense information helps further enhance the translation quality.

5 Conclusion

This work presents our approach to exploiting the WSD capabilities in LLMs to enhance the MT performance of sentences with ambiguous words. Specifically, we introduce two methods: "All Senses-based Prompting" and "One Predicted Sense-based Prompting", combined with two styles: *NLS* and *TS*. Experiments on the HOLLY test set highlight the effectiveness of our approach and underscore the importance of exploiting WSD capabilities in LLMs to improve MT.

Limitations

We evaluate our approach on a single benchmark dataset (the Korean-English HOLLY benchmark test set) since this dataset includes gold sense labels for homograph words (or ambiguous words) in the source sentences and provides the target sentences. However, we plan to test our approach on additional datasets as they become available in the future.

Ethics Statement

The linguistic expert, fluent in both Korean and English, helped to prepare three examples for few-shot learning, detailed further in Appendix A. They declined remuneration due to the minimal effort involved. Furthermore, as shown in Table 4, the three examples do not contain toxic content.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Jo hannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Yujin Baek, Ko tik Lee, Dayeon Ki, Hyoung-Gyu Lee, Cheonbok Park, and Jaegul Choo. 2023. Towards accurate translation via semantically appropriate application of lexical constraints. In *Annual Meeting of the Association for Computational Linguistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas Donald Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *ArXiv*, abs/2409.15790.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Rivière, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Akshay Uttama Nambi, Vaibhav Balloli, Mercy Prasanna Ranjit, Tanuja Ganu, Kabir Ahuja, Sunayana Sitaram, and Kalika Bali. 2023. Breaking language barriers with a leap: Learning strategies for polyglot llms. *ArXiv*, abs/2305.17740.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal S. Mian. 2023. A comprehensive overview of large language models. *ArXiv*, abs/2307.06435.

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–483, Nusa Dua, Bali. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, Patrick Haller, Renata

Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Van-Hien Tran, Chenchen Ding, Hideki Tanaka, and Masao Utiyama. 2023. Improving embedding transfer for low-resource machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 123–134, Macau SAR, China. Asia-Pacific Association for Machine Translation.

James Vo. 2024. Vi-mistral-x: Building a vietnamese language model with advanced continual pre-training. *ArXiv*, abs/2403.15470.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

A Appendix

The HOLLY Dataset. The HOLLY dataset (Baek et al., 2023) is a benchmark for evaluating Lexically-constrained Neural Machine Translation (LNMT) systems, focusing on handling homographs and lexical constraints in translation tasks. It assesses scenarios where lexical constraints are either semantically appropriate or not.

The dataset is divided into a training set, a validation set, and a test set. The training and validation sets are designed for a homograph disambiguation task and consist solely of Korean sentences. The training set contains 48,836 examples, while the validation set has 3,000 examples. Each example is a triplet of Korean sentences with a common homograph. The task is to determine if the homograph has the same meaning in all sentences (labeled "1") or if it differs in one (labeled "0").

The test set evaluates both homograph disambiguation and machine translation tasks, comprising 600 test examples. Each example in this test set includes a lexical constraint between a Korean homograph and its English meaning/sense, a source sentence with the homograph, and its English translation. Among these, 300 examples have correct lexical constraints (positive) and 300 have incorrect constraints (negative). The positive examples provide the gold sense label of the homograph, allowing evaluation of the sense prediction module as detailed in our "One Predicted Sense-based Prompting" method (Section 3).

Preparing for Few-Shot Learning. Here, we outline how a linguistic expert prepares three fixed examples for few-shot learning. This expert is fluent in both Korean and English. From the HOLLY training set, we randomly select three Korean source sentences, each containing one homograph word (ambiguous word). These homographs are unseen in the HOLLY test set.

The HOLLY training set, as mentioned earlier, includes only Korean source sentences without corresponding English target sentences. The linguistic expert's task involves identifying the correct sense of each homograph within its context, using the provided list of candidate senses. Once the correct sense is determined, the expert translates the entire source sentence into English.

Table 4 presents these examples in detail, showcasing the expert's translations. In our approach, described in Section 3, we use the first example for 1-shot learning scenario and all three examples for 3-shot learning scenario. Additionally, we explain the purpose of using the three samples with the linguistic expert.

Configurations of the ChrF++ Measure. Here are the configurations of the ChrF++ measure we used to evaluate MT quality. It uses a single reference translation ('nrefs:1'), is case-sensitive ('case:mixed'), and applies effective smoothing ('eff:yes'). The metric computes character n-gram precision and recall with 6-character n-grams ('nc:6') and 2-word n-grams ('nw:2'). Spaces are not considered as tokens ('space:no'). This configuration runs on version 2.4.1 of the chrF++ software, a tool designed to assess MT quality by comparing translations against reference texts.

id	Property	Content
1	Source Sent	한국에는 아파트나 빌라처럼 여러 가구가 살 수 있도록 지은 집이 많다.
	Target Sent	In Korea, there are many houses built to accommodate multiple households, such as apartments or villas.
	Homograph	가구
	All Senses	'household', 'furniture'
	Gold Sense	'household'
2	Source Sent	아버지의 사업 실패로 가산을 날려 민준이는 대학 등록금을 스스로 마련해야 했다.
	Target Sent	Due to the significant loss of the family fortune resulting from his father's business failure, Minjun had to finance his university tuition himself.
	Homograph	가산
	All Senses	'addition', 'family fortune'
	Gold Sense	'family fortune'
3	Source Sent	경찰은 일단 알리바이가 불명확한 사람이 범인이라는 가정을 세웠다.
	Target Sent	The police established the assumption that a person with an unclear alibi could be the culprit.
	Homograph	가정
	All Sense	'family', 'assumption'
	Gold Sense	'assumption'

Table 4: Three fixed examples for few-shot learning.