Can Uniform Meaning Representation Help GPT-4 Translate from Indigenous Languages?

Shira Wein
Amherst College
swein@amherst.edu

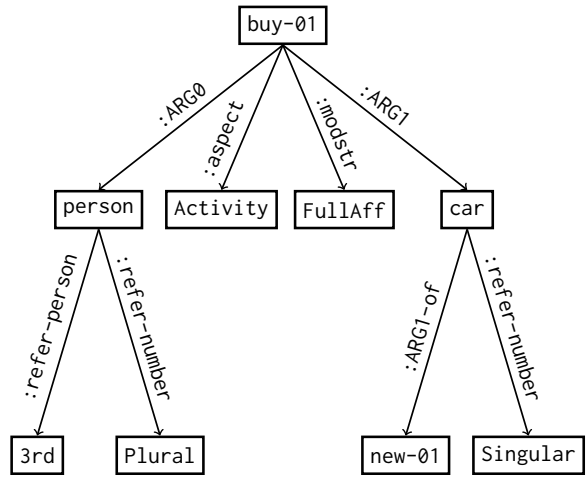
Abstract

While ChatGPT and GPT-based models are able to effectively perform many tasks without additional fine-tuning, they struggle with tasks related to extremely low-resource languages and indigenous languages. Uniform Meaning Representation (UMR), a semantic representation designed to capture the meaning of texts in many languages, is well-positioned to be leveraged in the development of low-resource language technologies. In this work, we explore the downstream utility of UMR for low-resource languages by incorporating it into GPT-4 prompts. Specifically, we examine the ability of GPT-4 to perform translation from three indigenous languages (Navajo, Arápaho, and Kukama), with and without demonstrations, as well as with and without UMR annotations. Ultimately, we find that in the majority of our test cases, integrating UMR into the prompt results in a statistically significant increase in performance, which is a promising indication of future applications of the UMR formalism.

1 Introduction

While ChatGPT models (Open AI, 2022) are able to successfully produce text in many highly-resourced languages, they severely struggle with machine translation of low-resource languages (Stap and Araabi, 2023; Robinson et al., 2023).

Uniform Meaning Representation (UMR; Van Gysel et al., 2021b) is a semantic representation created with the annotation of low-resource languages in mind. The UMR formalism is designed to represent a wide range of languages by providing flexibility in the annotation process via paradigmatic lattices and creating any required rolesets during "Stage 0" of the annotation process. UMR is a multilingual extension of the widely-adopted Abstract Meaning Representation (AMR; Banarescu et al., 2013).




```
(s / buy-01
  :ARG0 (p / person
    :refer-person 3rd
    :refer-number Plural)
  :ARG1 (c / car
    :ARG1-of (n / new-01)
    :refer-number Singular)
  :aspect Activity
  :modstr FullAff)
```

Figure 1: UMR graph for the sentence "They were buying a new car" in both graph form and in text-based 'PENMAN' notation (Kasper, 1989).

The first UMR dataset (Bonn et al., 2024) has recently been released, enabling exploration into the utility of UMR for tasks related to the generation of text into and from low-resource languages. Recent work has also shown that GPT models likely do not implicitly contain the linguistic knowledge necessary to construct an AMR graph (Ettinger et al., 2023)—or by extension, a UMR graph—suggesting that the addition of a UMR annotation may support prompt-based translation.

Thus, in this work, we explore the downstream benefits of incorporating UMR graphs into ChatGPT prompts, specifically with regard to machine translation from extremely low-resource languages into English. We craft four prompting protocols for

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 278–285

GPT-4[1] which vary in both their number of demonstrations and whether UMR is included: (1) zero-shot prompting, (2) zero-shot prompting with the UMR graph of the text included, (3) five-shot prompting, and (4) five-shot prompting with the UMR graphs included. We perform our experiments on three indigenous languages included in Bonn et al. (2024) which also contain English references: Navajo, Kukama, and Arápaho. Our contributions include:

- Prompting protocols for translating from indigenous languages, with and without demonstrations (i.e. zero- and five-shot), and with and without UMR graphs of the source text.
- Experiments producing English translations of more than 1000 individual source sentences across three extremely-low resource languages via GPT-4.
- Statistical analyses of the results of each of our protocols, which indicate the quantitative improvement that the incorporation of UMR graphs and demonstrations begets.

2 Background & Related Work

2.1 Machine Translation with ChatGPT

Recent work has explored the effectiveness of prompting GPT models to generate text in no- and low-resource languages, with generally poor indications of success (Stap and Araabi, 2023).

Guo et al. (2024) focus on mitigating the issue of data sparsity for low-resource translation via ChatGPT and BLOOMZ (Muennighoff et al., 2023) by providing a vocabulary list and demonstrations as additional input.

Notably, Robinson et al. (2023) find that ChatGPT performs competitively with state-of-the-art machine translation models for high-resource languages, but performs poorly for low-resource languages. In particular, the most significant predictor of ChatGPT translation performance on a language is the number of Wikipedia entries that exist in the language, serving as a proxy of how well-sourced that language is. Additionally, five-shot prompts lead to small performance gains over zero-shot prompts. Tang et al. (2025) further indicate that (in a high-resource setting) adaptive few-shot prompting, which uses the most semantically similar texts in the dataset to the source text as demonstrations, leads to increased performance gains.

[1] https://openai.com/index/gpt-4/

Related work has explored the utility of chain-of-thought prompting for translating with ChatGPT, finding it to be generally ineffective as it results in word-by-word translation (Peng et al., 2023).

2.2 Uniform Meaning Representation

Uniform Meaning Representation (UMR) is an extension of the popular semantic representation Abstract Meaning Representation (AMR). AMR is a graph-based semantic representation which captures "who does what to whom," reflecting the semantic relationships within the sentence. The nodes in an AMR graph correspond with concepts in the sentence (or phrase), while edges denote the relationships between those concepts. UMR, like AMR, represents the relationships between concepts in a sentence in the form of a rooted, directed graph (see Figure 1 for an example UMR in both text-based and graph-based form).

While AMR was originally designed for English (Banarescu et al., 2013), UMR is designed to be multilingual and contains information about the text at both the sentence- and document-level (Van Gysel et al., 2021b). UMR accommodates a range of linguistic features in comparison to AMR (Wein and Bonn, 2023) through the integration of lattice-based annotation structures, which allow the annotator to select the level of granularity appropriate for the individual language (Van Gysel et al., 2019). UMR is also particularly well-suited to the annotation of low- and no-resource languages, including indigenous languages (Van Gysel et al., 2021a), as it incorporates dataset development (in the form of rolesets) into "Stage 0" of the annotation process (Vigus et al., 2020), thus overcoming the lack of preexisting rolesets for some languages. UMR's design, which both accommodates linguistic diversity and overcomes a lack of data for low-resource languages, motivates its use in our work.

While AMR is a semantic representation that has been widely adopted and has proven useful in many monolingual settings (Wein and Opitz, 2024), it is not possible to annotate low-resource languages in AMR. A language-specific version of AMR would need to be created for each individual low-resource language, as there is not sufficient annotation flexibility in AMR to effectively annotate multilingually (Wein and Schneider, 2024), thus making AMR not appropriate for our work. Still, prior work has demonstrated that AMR is particularly useful in low-resource engineering studies (Hua et al., 2023; Gururaja et al., 2023; Ghosh et al., 2024), which

motivates incorporating UMR into low-resource settings and for low-resource languages.

3 Methodology

3.1 Data

In this work, we examine the utility of incorporating UMR graphs into GPT-4 prompts which instruct the system to translate a source text in the extremely low-resource languages of Navajo, Kukama, and Arápaho into English.

The recently released UMR dataset which is leveraged in this work (Bonn et al., 2024) contains sentences from English (209 sentence-level graphs, 202 document-level), Chinese (358 sentence-level graphs, 358 document-level), Arápaho (406 sentence-level graphs, 109 document-level), Navajo (506 sentence-level graphs, 168 document-level), Kukama (105 sentence-level graphs, 86 document-level), and Sanapaná (602 sentence-level graphs, 602 document-level). Arápaho, Navajo, Kukama, and Sanapaná are all indigenous languages which are extremely low-resource. Not all annotations contain both sentence-level and document-level graphs. The Navajo, Kukama, and Arápaho UMR graphs all provided English translations with the annotations, while the Sanapaná annotations contained Spanish translations. While included in the UMR dataset, we forgo translation from Sanapaná, as English translations are not provided, negating our ability to use English texts as references when evaluating the system output.

We generate translations from the 506 sentences in Navajo, 105 sentences in Kukama, and 406 sentences in Arápaho, resulting in 1,017 total sentences translated.

3.2 Prompts

We design and use four prompting protocols:

1. Zero-shot: Instruct the model to translate from the source language into English, providing the text to be translated.
2. Zero-shot with UMR: Instruct the model to translate from the source language into English, providing the text to be translated and the UMR of the source text.
3. Five-shot: Instruct the model to translate from the source language into English, providing five demonstrations (texts in the source language as well as their reference English translations), as well as the text to be translated.
4. Five-shot with UMR: Instruct the model to translate from the source language into English, providing five demonstrations (texts in the source language as well as their reference English translations, plus their UMRs), the text to be translated, and the UMR of the source text.

For our five-shot prompts, we use an adaptive approach to demonstration selection, selecting the 5-nearest neighbors to the source sentence as the demonstrations. We use chrF to compare the source language text to the other sentences in that language. We are using the source language sentences (rather than the English references) to identify the five most relevant demonstrations, to ensure that the same would be possible at test time.

The specific text contained in each prompt can be seen in Figure 2 in Appendix A.

3.3 Evaluation

We perform translation *from* the indigenous languages *into* English in order to enable more accurate evaluation of the generated text, via automatic and qualitative analyses.

We evaluate the performance of the model for each item using each of the four prompting protocols.[2] The automatic metrics we use to evaluate our generated text are chrF (Popović, 2015) and BERTscore (Zhang et al., 2020).

4 Results

Table 1 shows the average BERTscore and chrF values, respectively, of each prompting protocol in each language. While the BERTscore values are all closer together (as expected, given that BERTscore struggles to capture finer-grained differences in meaning (Leung et al., 2022; Wein et al., 2023)), we can see generally that for BERTscore, the five-shot with UMR scores are highest, followed by five-shot scores. Then, with a notable decrease in BERTscore value, the zero-shot and zero-shot with UMR scores follow, but switching which of the two is higher. For the chrF scores, the five-shot with UMR scores are highest for all languages, followed by the five-shot scores, next followed by the zero-shot with UMR scores (again notably lower than the five-shot performance) and finally the zero-shot scores.

While these average scores provide an initial impression as to the benefits of adding UMR graphs

[2]This experimentation cost $62.11 USD in OpenAI credits.

Evaluation Metric	Prompting Protocol	Arápaho	Kukama	Navajo
BERTscore	Zero-Shot	0.867±0.02	0.862±0.02	0.862±0.02
	Zero-Shot w UMR	0.867±0.05	0.857±0.03	0.867±0.03
	Five-Shot	0.903±0.04	0.904±0.04	0.885±0.03
	Five-Shot w UMR	0.910±0.04	0.912±0.04	0.891±0.03
chrF	Zero-Shot	13.0±5.5	14.0±5.8	15.4±6.4
	Zero-Shot w UMR	16.2±8.7	16.8±7.0	17.9±8.3
	Five-Shot	32.9±21	40.8±25	24.6±14.2
	Five-Shot w UMR	35.7±22	43.5±24	25.9±14.1

Table 1: Average scores for each language, prompting protocol, and evaluation metric. Standard deviation is indicated after the plus or minus sign.

	Arápaho	Kukama	Navajo
BERTScore: Zero-shot vs Zero-shot with UMR	0.9721	0.0146	**<0.0001**
chrF: Zero-shot vs Zero-shot with UMR	**<0.0001**	**<0.0001**	**<0.0001**
BERTScore: Zero-shot vs Five-shot	**<0.0001**	**<0.0001**	**<0.0001**
chrF: Zero-shot vs Five-shot	**<0.0001**	**<0.0001**	**<0.0001**
BERTScore: Five-shot vs Five-shot with UMR	**<0.0001**	**0.0017**	**<0.0001**
chrF: Five-shot vs Five-shot with UMR	**0.0004**	0.0555	**0.0294**

Table 2: Two-tailed paired t-test p-values for statistical comparisons of the BERTscores and chrF scores for each prompting protocol, in each language. The bolded entries indicate a statistically significant improvement when adding UMR or demonstrations. The red entry (first row of Kukama scores) highlights a statistically significant difference, but where the zero-shot *without* UMR performs better than the zero-shot *with* UMR.

and demonstrations in the prompt, we perform two-tailed paired t-tests comparing the automatic metric scores of the output from the various prompts in order to ascertain whether the inclusion of a UMR graph and/or the five demonstrations results in a statistically significant increase in the score (and therefore, performance). Specifically, we compare the scores for (1) zero-shot and zero-shot with UMR prompts, (2) zero-shot and five-shot prompts, and (3) five-shot and five-shot with UMR prompts. These values can be seen in Table 2.

We find that for 9 of the 12 comparisons of prompts with UMR versus without UMR, adding the UMR to the prompt results in a statistically significant increase. For two comparisons (Arápaho BERTscore: Zero-shot vs Zero-shot with UMR; Kukama chrF: Five-shot versus Five-shot with UMR), there is no statistical difference, and in one case (Kukama BERTscore zero-shot vs zeros-show with UMR), there is a statistically significant difference but the scores from the prompt without UMR are higher. These findings indicate that the inclusion of a UMR graph into a prompt enables more effective text generation in extremely low-resource settings, as the UMR may be supplying additional linguistic information not already contained in the model.

Additionally, for all 6 of the cases where we compare zero-shot scores against five-shot scores, there is a statistically significant improvement. This indicates that the use of demonstrations is also useful for prompting with extremely low-resource languages. While both UMR and demonstrations lead to improvements in translation quality, the most drastic difference is found when adding the demonstrations to move from zero-shot to five-shot prompting.

As indicated by our quantitative results and statistical analysis, our qualitative analysis further suggests that the English translation more closely resembles the reference when incorporating UMR and/or demonstrations into the prompt.

Take as an example the following Kukama item, which has a disfluent English reference of "He run in the forest:" *ay ra yupuni yapana ɨwɨrati*. The zero-shot translation is completely unrelated to the text, though it does have some reference to a man performing an action in a natural setting: "He plays with his younger brother at the river." The zero-shot with UMR text is again unrelated, indicating a person performing an action in an unspecified setting: "The person is working there today." Then, the five-shot text contains more semantic similarity with the reference, as there is a male moving in the forest: "He has already started walking in the forest." Finally, the five-shot text with UMR shows the male to be running in the forest, clearly exhibiting the most semantic similarity with the reference: "He has already started running in the forest." This example exemplifies the fact that five-shot prompting

alone is not enough to achieve optimal performance on this task, and leveraging UMR in the prompt benefits performance beyond what is possible only when using demonstrations.

Another example is the Arápaho sentence, *wohei noh ci'ceese' hoo3itoo, heetnoo3itoone3en*, which has an English reference of "Wohei another story, I'm going to tell you another story." The zero-shot text is completely unrelated to the reference: "I walked to the store and said hello to the shopkeeper." The zero-shot with UMR output, on the other hand, is much more semantically similar to the reference: "I will tell you a story." The five-shot text then contains some of the specific language included in the reference, as well as some semantic similarity: "Wohei, then go ask him, the storyteller." Finally, the five-shot with UMR text is the most semantically similar, though the ending is slightly harder to interpret: "Wohei I will tell you a story about a little."

Our automatic metrics and qualitative analyses reveal that, on our test data, for most cases, incorporation of UMR graphs and demonstrations into the prompts enables heightened similarity with the reference English translation. Therefore, leveraging UMR in the prompt does indeed lead to heightened performance when translating from indigenous languages into English, while the greatest improvements are achieved by using related sentences as demonstrations; the combination of using both the demonstrations and the UMR in the prompt leads to the highest quality output in our experiments.

5 Conclusion

In this work, we begin to address one of the failings of GPT-based models: that of translation from extremely low-resource languages. We specifically examine the ability of a newly released Uniform Meaning Representation (UMR) dataset—containing sentences in Navajo, Arápaho, and Kukama, their UMR graphs, and their parallel sentences in English—to improve GPT-4 performance when included in the prompt. We find that both the incorporation of UMR graphs of the source text and adaptively selected demonstrations lead to improved performance on low-resource machine translation via prompting, with a statistically significant increase resulting in the majority of our comparisons. This is a promising indication of the downstream utility of UMR for low-resource settings and a step forward towards effective translation from indigenous languages via prompting.

Limitations

We perform experimentation on three extremely low-resource indigenous languages. Future work could expand this evaluation to other languages as well, varying in their depth of resources, as additional UMR annotations are released. UMR annotation can be expensive and time-consuming, as it requires fluency in the language and annotator training, which is a barrier to seamlessly incorporating UMR into downstream applications.

Additionally, we perform translation in one direction, with the low-resource languages serving as the source and English serving as the target language. Performing translation from English into these low-resource languages would make for interesting future work, though it will require a human evaluation by speakers of the language.

Finally, randomness is inherent in the results generated from GPT-4. We attempt to curtail this effect by providing statistical analyses for our findings, but further rigor could be added by running these experiments additional times.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, Jens E. L. Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italy. ELRA and ICCL.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "You are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.

Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, Chandra Kiran Evuru, Ramaneswaran S, S Sakshi, and Dinesh Manocha. 2024. ABEX: Data augmentation for low-resource NLU via expanding abstract descriptions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 726–748, Bangkok, Thailand. Association for Computational Linguistics.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Sireesh Gururaja, Ritam Dutt, Tinglong Liao, and Carolyn Rosé. 2023. Linguistic representations for fewer-shot relation extraction across domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7514, Toronto, Canada. Association for Computational Linguistics.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada. Association for Computational Linguistics.

Robert T. Kasper. 1989. A flexible interface for linking applications to Penman's sentence generator. In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.

Wai Ching Leung, Shira Wein, and Nathan Schneider. 2022. Semantic similarity as a window into vector- and graph-based metrics. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 106–115, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Open AI. 2022. Introducing ChatGPT.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Lei Tang, Jinghui Qin, Wenxuan Ye, Hao Tan, and Zhijing Yang. 2025. Adaptive few-shot prompting for machine translation with pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25255–25263.

Jens E. L. Van Gysel, Meagan Vigus, Lukas Denk, Andrew Cowell, Rosa Vallejos, Tim O'Gorman, and William Croft. 2021a. Theoretical and practical issues in the semantic annotation of four indigenous languages. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 12–22, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jens E. L. Van Gysel, Meagan Vigus, Pavlina Kalm, Sook-kyung Lee, Michael Regan, and William Croft. 2019. Cross-linguistic semantic annotation: Reconciling the language-specific and the universal. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 1–14, Florence, Italy. Association for Computational Linguistics.

Jens E.L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021b. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

Meagan Vigus, Jens E. L. Van Gysel, Tim O'Gorman, Andrew Cowell, Rosa Vallejos, and William Croft. 2020. Cross-lingual annotation: a road map for low- and no-resource languages. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 30–40, Barcelona Spain (online). Association for Computational Linguistics.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 23–33, Nancy, France. Association for Computational Linguistics.

Shira Wein and Juri Opitz. 2024. A survey of AMR applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6856–6875, Miami, Florida, USA. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2024. Assessing the cross-linguistic utility of Abstract Meaning Representation. *Computational Linguistics*, 50(2):419–473.

Shira Wein, Zhuxin Wang, and Nathan Schneider. 2023. Measuring fine-grained semantic equivalence with Abstract Meaning Representation. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 144–154, Nancy, France. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proc. of ICLR*, Online.

A Prompting Protocols

Zero-shot

```
Please provide the English translation for this [Source language] sentence. Do not provide
any explanations or text apart from the translation.
[Source Language]: [sentence to be translated]
English:
```

Zero-shot with UMR

```
Please provide the English translation for this [Source language] sentence (which is
accompanied by a Uniform Meaning Representation parse). Do not provide any explanations or text apart
from the translation.
[Source language]: [sentence to be translated]
Uniform Meaning Representation: [UMR of source text]
English:
```

Five-shot

```
Please provide the English translation for this [Source language] sentence. Do not provide
any explanations or text apart from the translation.
[Source language]: [sentence 1] English: [translation 1]
[Source language]: [sentence 2] English: [translation 2]
[Source language]: [sentence 3] English: [translation 3]
[Source language]: [sentence 4] English: [translation 4]
[Source language]: [sentence 5] English: [translation 5]
Please provide the English translation for this [Source language] sentence.
Do not provide any explanations or text apart from the translation.
[Source language]: [sentence to be translated]
English:
```

Five-shot with UMR

```
Please provide the English translation for this [Source language] sentence (which is
accompanied by a Uniform Meaning Representation parse). Do not provide any explanations or text apart
from the translation.
[Source language]: [sentence 1] Uniform Meaning Representation: [UMR 1] English: [translation 1]
[Source language]: [sentence 2] Uniform Meaning Representation: [UMR 2] English: [translation 2]
[Source language]: [sentence 3] Uniform Meaning Representation: [UMR 3] English: [translation 3]
[Source language]: [sentence 4] Uniform Meaning Representation: [UMR 4] English: [translation 4]
[Source language]: [sentence 5] Uniform Meaning Representation: [UMR 5] English: [translation 5]
Please provide the English translation for this [Source language] sentence (which is accompanied by
a Uniform Meaning Representation parse). Do not provide any explanations or text apart from the
translation.
[Source language]: [sentence to be translated]
Uniform Meaning Representation: [UMR of source text]
English:
```

Figure 2: The "user" portions of our prompts for the four protocols. For all prompts, the "system" portion of the protocol is as follows: `System: You are a machine translation system from [Source language] to English that translates sentences from narrative documents.`