UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D17 – First report on adaptable and context

Nature	Report	**Work Package**	WP4
Due Date	28/03/2024	**Submission Date**	dd/mm/2024
Main authors	Chrysoula Zerva (IT)		
Co-authors	Barry Haddow (UEDIN), Alexandra Birch (UEDIN) Mateusz Klimaszewski (UEDIN), Pinzhen Chen (UEDIN), José Souza (Unbabel)		
Reviewers	Barry Haddow and Alexandra Birch		
Keywords	adaptation, prompting, machine translation, context-awareness, speech translation		
Version Control			
v0.1	**Status**	Draft	26/03/24
v1.0	**Status**	Final	28/03/24

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

List of Figures

Abstract

In this deliverable, we report research updates for WP4, carried out within the first half of the UTTER project. Our focus within this work package is on developing methods to better adapt pretrained neural models to generation and translation tasks. Beyond developing and evaluating better adaptation techniques, we put emphasis on contextualization methods to better tune and personalise model outputs. Finally, we start exploring simultaneous translation aiming to focus more on this task in the second half of the project. Overall, we have seen great advances in the first half of the project and do not anticipate any noteworthy risk, expecting steady progress in the second half of the project.

1 Introduction

In WP4 we aim to focus on the development of accurate generation and translation models for spoken and textual dialogue. We aim to do so by tuning the pretrained models developed in WP3 on the multilingual dialogue data created in WP2, developing:

- Dynamic adaptation and contextualization techniques, that boost performance on generation and machine translation tasks.
- Controlled generation techniques to incorporate the speakers' preferences (e.g. formality, gender) and predictive models of speakers' emotional state, towards more empathic conversations.
- Cascaded and end-to-end approaches for translating spoken language, with a focus on data and model efficiency, robustness, translation quality, and evaluation.
- New techniques for simultaneous translation.

Our work in the first half of the project led to a total of:

- Six **co-organised shared tasks**: (a) the *WMT Chat Translation Shared Task* in 2022; (b) the *WMT Quality Estimation Shared Tasks* in 2022 and 2023; (c) the *WMT Metrics Shared Tasks* in 2022 and 2023; and (d) the IWSTL Shared Task in 2023. We are also co-organising three shared tasks in the upcoming year (WMT Chat Translation Shared Task 2024[1]; Quality Estimation Shared Task 2024; and the Metrics Shared Task 2024).
- 29 **manuscripts** of which:
  - 1 journal paper (TACL)
  - 23 in conferences (8 of them ranked A*, 4 ranked A): WMT 2022 (5 papers), EACL 2023 (3 papers), ICLR 2023 (1 paper), ACL 2023 (1 paper), IWSLT 2023 (1 paper), EMNLP 2023 (4 papers), WMT 2023 (5 papers), EACL 2024 (1 paper), LREC-COLING 2024 (1 paper)
  - 5 arXiv pre-prints
- 15 repositories releasing code and data:
  - https://github.com/Unbabel/COMET/
  - https://github.com/Vicky-Wil/ReMaKE
  - https://github.com/deep-spin/tower-eval
  - https://github.com/deep-spin/tower-alignment
  - https://github.com/deep-spin/efficient_kNN_MT
  - https://github.com/deep-spin/translation-hypothesis-ensembling
  - https://github.com/deep-spin/translation_llm?
  - https://github.com/CoderPat/MuDA
  - https://github.com/Wafaa014/context-utilization
  - https://github.com/su0315/discourse_context_mt

[1] https://www2.statmt.org/wmt24/chat-task.html

- https://github.com/Remorax/CCS-Pretraining-NMT
- github.com/johndmendonca/MAIA-DQE
- https://github.com/WMT-QE-Task/wmt-qe-2023-data
- https://github.com/WMT-QE-Task/wmt-qe-2022-data
- https://github.com/WMT-Chat-task/

2 Task 4.1: : Adaptable, multimodal generation and translation (IT*, UEDIN, UVA, NAV)

Proposal highlights

In this task we focus on adaptation strategies for multiple generation tasks: machine translation (MT), transcription, and summarization. The key proposal highlights are listed below.

- Dynamic adaptation techniques that combine in-context learning and adaptors as an alternative to expensive fine-tuning.
- Investigating new adaptation techniques that can handle multiple modalities.
- Direct and cascaded approaches to speech translation, using large language models.

Summary of completed work

Since the kickoff of the UTTER project we have seen an unprecedented advancement of large language models (LLMs) and related technologies (Min et al., 2023), which further emphasised the importance of exploring novel adaptation methodologies for generation and translation. Hence a significant amount of work in the first half of the project focused on adaptation methodologies that allow better steering of large language models to translation or other downstream tasks, ranging from prompting to instruction tuning approaches (see Section 2.1. In addition, striving for robust and highly multilingual models, in Section 2.2 we describe works that propose adaptation methodologies that focus on performance improvements for low-resource languages and multilingual settings. Sections 2.3 and 2.4 describe works that focused on extending the capabilities of generation models, either by extending their knowledge beyond the training data via the proposal of novel retrieval-augmented methodologies or trying to analyse and mitigate specific challenging phenomena in MT. Finally, we report a set of works on tuning models for machine translation evaluation in Section 2.5.

2.1 Steering Large Language Models

The works reported in this section focus on adaptation methods for LLMs, that target specific generation tasks, with emphasis on improving machine translation. Specifically, our work focused on either (a) exploring different prompting and in-context learning scenarios to harness the capabilities of LLMs and improve performance for a given task without further tuning, and (b) investigation and advancement of instruction-tuning methodologies to better adapt to MT tasks. The work in §2.1.5 takes a step further, analysing the potential of ensembling methods for MT with LLMs, comparing their impact when used in prompting, instruction-tuning as well as minimum Bayes

risk (MBR) decoding (Eikema and Aziz, 2020). Finally, the survey presented in §2.1.6 paves the path for future alignment-related methodologies to control LLM perfomance.

2.1.1 Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting

Terminology correctness is important in the downstream application of machine translation, e.g. "apple" should sometimes be translated as a fruit and sometimes as a named entity depending on the context and user preference. Our research looked at two approaches to terminology-aware translation—adapting machine translation models to produce desired terminology terms, which are usually supplied by the user in the format of pairs of source and target words. Targeting traditional neural machine training and decoding, we incorporated training data with pseudo-terminology labels produced by word alignment tools, ran inference with real-time terminology labels, and also experimented with negatively constrained decoding to "fix" output translations that do not obey pre-defined terminology constraints. Moving to large language models, we investigate the idea of feeding terminology constraints as natural texts via prompting, following previous works that target a more general-purpose translation pipeline.

We participated in the 2023 shared task on terminology translation with our algorithms. Our translation submissions are rated as high-quality and consistent in terminology injection (Semenov et al., 2023).

A full description of this work can be found in our paper (Bogoychev and Chen, 2023).

2.1.2 Prompting large language model for machine translation: A case study

Large language models (LLMs) pretrained on massive unlabeled corpora have shown impressive emergent abilities under model scaling which enable prompting for downstream applications (Brown et al., 2020; Kaplan et al., 2020). Different from task-specific finetuning, prompting constructs task-specific prompts by rephrasing test examples with descriptive task instructions and executes the task by feeding prompts to LLMs directly. It can be further enhanced through in-context learning by providing a few labelled examples (or prompt examples) as a demonstration (Brown et al., 2020). As a new paradigm, prompting LLMs has achieved state-of-the-art performance over a range of natural language processing (NLP) tasks (Chung et al., 2022).

In this paper, we focus on prompting LLMs for machine translation (MT). MT represents a complex task requiring transforming a source input into its semantically equivalent target output in a different language, which combines sequence understanding and generation. It offers a unique platform to assess the cross-lingual generation capability of LLMs, and the assessment may shed light on pretraining/finetuning algorithm design for achieving universal LLMs (Chowdhery et al., 2022). While a few studies have reported translation results (Brown et al., 2020; Chowdhery et al., 2022), a systematic study on how prompting works for MT is still missing in the literature.

We aim at filling this gap by thoroughly examining different prompting setups using the recently released GLM (Zeng et al., 2022), particularly concerning three aspects: *the prompting strategy*, *the use of unlabeled/monolingual data*, and *the feasibility of transfer learning*. Prompting has shown varying sensitivity to the choice of prompt templates and examples. For MT, prior studies adopted different templates (Brown et al., 2020; Chowdhery et al., 2022), and we reevaluate them to figure out the optimal one. We further design a set of features for prompt examples and ex-

plore which one(s) could explain the prompting performance, according to which we develop the example selection strategy.

Since leveraging monolingual data to improve MT has long been of interest, we would like to determine whether and how such data can be used in prompt example construction. We make a step in this direction by studying the effect of data augmentation using back-/forward-translation (Sennrich et al., 2016a) via zero-shot prompting. In addition, neural MT and pretrained LLMs have shown encouraging transfer abilities (Devlin et al., 2019) but transfer learning for prompting has received little attention. Whether prompt examples are transferable across different settings, such as from one domain/language pair to another and from sentence-level examples to document-level translation, is yet to be addressed.

We address the above concerns with GLM as the testbed and conduct extensive experiments on Flores and WMT evaluation sets. We mainly study translation for three languages: English, German and Chinese. We also provide a quantitative and qualitative analysis to disclose problems when prompting for MT, which might offer insights for future study. Our main findings are listed as below:

- Prompting performance varies greatly across templates, and language-specific templates mainly work when translating into languages LLMs are pretrained on. An English template in a simple form works best for MT.
- Several features of prompt examples, such as sequence length, language model score, and semantic similarity, correlate significantly with its prompting performance while the correlation strength is weak in general. Selecting examples based on these features can outperform the random strategy, but not consistently.
- Using monolingual examples for prompting hurts translation. By contrast, constructing pseudo-parallel examples via back-/forward-translation is a good option. Back-translation performs better and is more robust.
- Prompting shows some degree of transferability. Using demonstrations from other settings can improve translation over the zero-shot counterpart, while the superiority of a demonstration in one setting can barely generalize to another.
- Prompting for MT still suffers from copying, mistranslation of entities, hallucination, inferior direct non-English translation, and prompt trap where translating the prompt itself via prompting becomes non-trivial.

A full description of this work can be found in our paper (Zhang et al., 2023).

2.1.3 Steering Large Language Models towards Translation

Large language models (LLMs) are a promising avenue for translation through in-context learning. While this approach avoids supervision on parallel data, its effectiveness is example-dependent. When parallel data is available, LLMs can be finetuned on translation instructions, outperforming few-shot prompting and eliminating the need for in-context examples. However, traditional finetuning incurs in a high training cost and it remains unclear whether finetuned LLMs still benefit from desirable in-context learning properties, such as domain adaptation. In this work, we

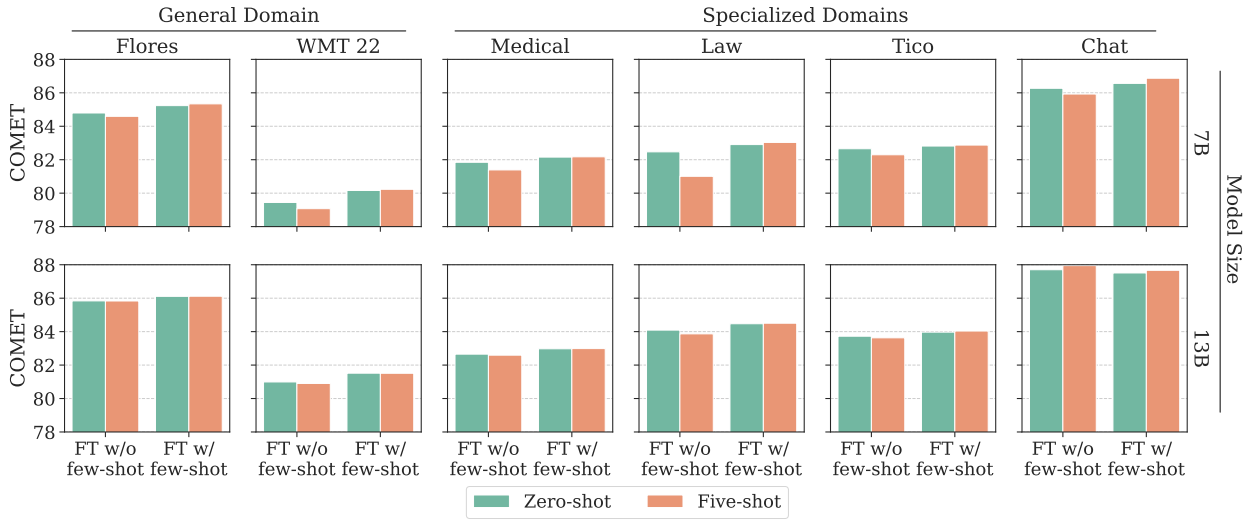

Figure 1: Results for zero-shot and five-shot translation by finetuning on translation instructions with and without in-context examples.

examine the impact of finetuning and few-shot prompting to adapt LLMs for translation. We show that LoRA achieves similar translation quality to full finetuning at a lower training cost, both outperforming few-shot prompting. However, we also find that finetuning on translation instructions degrades in-context learning capabilities. We address this issue by introducing in-context examples during finetuning, recovering few-shot capabilities while maintaining the benefits of finetuning — see Figure 1.

A full description of this work can be found in our paper (Alves et al., 2023).

2.1.4 TowerLLM

Many important tasks within multilingual NLP, such as quality estimation, automatic post-edition or grammatical error correction, are relevant to translation workflows — we call these translation-related tasks. While general-purpose large language models (LLMs) demonstrate proficiency on multiple tasks within the domain of translation, approaches based on open LLMs are competitive only when specializing on a single task. We developed an open LLM specialized for translation-related tasks through a three-step process. First, we extend LLaMA-2's multilingual capabilities with continued pretraining on a highly multilingual corpus — see report D3.1 for further details. Second, we create a dataset to specialize LLMs for translation-related tasks. Third, we perform supervised finetuning to obtain an instruction-following model tailored for translation-oriented and related multilingual and cross-lingual tasks.

We put together a dataset with relevant tasks to translation workflows, applied before or after translation the translation step. In building this dataset, we prioritize data diversity, collecting records from existing datasets spanning various domains for each task, and reformulating them as question-answer pairs using multiple manually curated templates. Additionally, we increase task diversity with the inclusion of paraphrasing, dialogue data and coding instructions. We also focus on data quality, constructing our question-answer pairs from human-annotated records and employing multiple quality filters. Our specialized model, TowerInstruct 13B consistently achieves higher translation quality than other open alternatives and is competitive with the closed GPT-4

and GPT-3.5-turbo models — see Table 1. Additionally, it outperforms open models on automatic post-editing, grammatical error correction and named entity recognition — see Table 2.

Models	Flores-200 en→xx	Flores-200 xx→en	WMT 23 en→xx	WMT 23 xx→en	TICO 19 en→xx
Closed					
GPT-3.5-turbo	88.95 2	88.14 3	85.56 2	83.48 2	87.36 2
GPT-4	**89.13** 1	**88.42** 1	**86.01** 1	**83.69** 1	**87.52** 1
Open					
NLLB 54B	86.79 4	87.95 3	78.60 7	79.06 6	87.05 2
LLaMA-2 70B	87.82 4	88.19 2	82.95 6	82.56 4	86.46 4
Mixtral-8x7B-Instruct	87.76 3	88.17 2	83.60 5	82.84 3	86.60 4
ALMA-R 7B	—	—	83.40 5	82.39 4	—
ALMA-R 13B	—	—	84.46 3	83.03 3	—
TOWERINSTRUCT 7B	88.51 3	88.27 2	84.28 3	82.77 4	87.01 3
TOWERINSTRUCT 13B	88.88 2	**88.47** 1	85.14 2	83.18 2	87.32 2

Table 1: Results for machine translation aggregated by language pair. Models with statistically significant performance improvements are grouped in quality clusters. We highlight the best ranked models in bold and underline the best ranked open models.

Models	APE↑ en→xx	APE↑ xx→en	GEC↓ Multilingual	NER↑ Multilingual
Baseline (no edits)	76.80	79.99	16.66	—
Closed				
GPT-3.5-turbo	81.47 4	78.68 5	15.06 2	50.22 4
GPT-4	**85.20** 1	**84.30** 1	15.08 2	59.88 3
Open				
LLaMA-2 70B	78.34 5	81.03 4	21.74 5	44.62 5
Mixtral-8x7B-Instruct	82.64 3	82.81 2	17.10 4	41.77 6
TOWERINSTRUCT 7B	82.69 2	81.56 4	15.13 3	71.68 2
TOWERINSTRUCT 13B	83.31 2	82.26 2	15.68 2	**74.70** 1

Table 2: Results for translation-related tasks aggregated by language or language pair. Models with statistically significant performance improvements are grouped in quality clusters. We highlight the best ranked models in bold and underline the best ranked *open* models. Since GEC is a held out task, we evaluate all models with 5 in-context examples.

A full description of this work can be found in our paper (Alves et al., 2024). The Tower models are available in the Tower HuggingFace collection[2].

[2] https://huggingface.co/collections/Unbabel/tower-659eaedfe36e6dd29eb1805c

2.1.5 Translation Hypothesis Ensembling with Large Language Models

Large language models (LLMs) are becoming a one-fits-many solution, but they sometimes hallucinate or produce unreliable output (Guerreiro et al., 2023). In this paper, we focus on the specific task of machine translation and investigate how hypothesis ensembling can improve the quality of the generated text. We experiment with several techniques for ensembling hypotheses produced by LLMs such as ChatGPT[3], LLaMA (Touvron et al., 2023), and Alpaca (Taori et al., 2023), providing a comprehensive study along multiple dimensions: the method to generate hypotheses (multiple prompts, temperature-based sampling, and beam search) and the strategy to produce the final translation (instruction-based, quality-based reranking, and minimum Bayes risk (MBR) decoding).

Our main findings can be summarized as follows. First, we demonstrate that translation quality can be enhanced with a small number of samples (*e.g.*, 20), especially when translating out of English. Notably, this differs from the findings of previous research using task-specific NMT models (Fernandes et al., 2022; Freitag et al., 2022a). Second, we discuss in which conditions beam search remains a reliable baseline for single-hypothesis translation and how to ensemble translations. Moreover, we find that there exists a significant gap in the quality of ensembles of unbiased samples from LLaMA and Alpaca. We attribute this disparity to how instruction tuning affects the relationship between the diversity of the hypotheses and the sampling temperature, which ultimately impacts translation quality. Lastly, we show that hypothesis ensembling reduces the number of generated hallucinations, thereby improving the model's robustness to source perturbations. Ensembling predictions and increasing the model size narrows the quality gap between open-source models and ChatGPT.

A full description of this work can be found in our paper (Farinhas et al., 2023).

2.1.6 Bridging the gap: A survey on integrating (human) feedback for natural language generation

In this survey, we explore the role of human feedback in advancing and "aligning" natural language generation (NLG) systems. We discuss problems with state-of-the-art language models, such as generating toxic or inaccurate content, and how **human feedback** on their outputs might solve these problems.

We specifically begin by providing a formalization of feedback, categorizing feedback types, discussing its formats and objectives, and finally, we outline direct and indirect methods for incorporating feedback into model training and improvement. We also discuss potential issues around the collection of feedback. Finally, we also account for AI feedback, which leverages powerful LLMs to minimize the need for human feedback. The main works covered can be seen in the taxonomy diagram of Figure 2.

We also trace possible future directions in the field, and we summarize the main pathways below.

- Current models often underutilize more expressive forms of feedback like natural language, favouring ranking-based or numerical feedback. Methods that manage to effectively explore more expressive feedback would be an important path for future research.

[3] https://chat.openai.com/

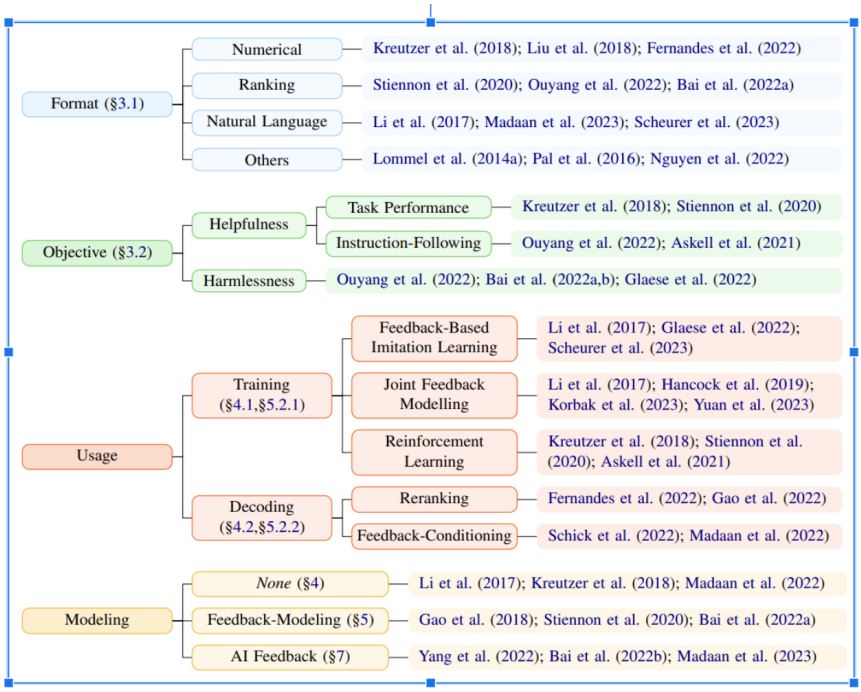

Figure 2: Taxonomy of methods that leverage human feedback, with some example representative works.

- A trade-off exists between the effort spent creating datasets and the reliability of judgments. Enlisting expert and diverse annotators can be beneficial and even crucial for high-stakes applications.
- The value of leveraging feedback lies primarily in the feedback itself rather than the specific method. While Reinforcement Learning from Human Feedback (RLHF) has been popular, other methods report notable improvements and might be simpler to apply (Gao et al., 2023; Rafailov et al., 2024; Zhou et al., 2024a; Zhao et al., 2023). However, large-scale comparative analysis remains necessary.

A full description of this work can be found in our paper (Fernandes et al., 2023a).

2.2 Adaptation to multilingual settings

When adapting neural and large language models to machine translation tasks, there is always a risk of favouring specific high-resource languages to the detriment of low-resource and especially non-English-centric ones. The works presented in this section analyse different aspects of this phenomena and propose methodologies to improve generation performance in highly multilingual settings.

2.2.1 When does monolingual data help multilingual translation: The role of domain and model scale

Recent work (Siddhant et al., 2022; Bapna et al., 2022; NLLB team et al., 2022) has shown that multilingual machine translation (MMT) using large amounts of parallel and monolingual data is

an effective way of improving MT performance for low-resource language pairs. Typically the monolingual data is incorporated using a combination of denoising autoencoding (DAE; Conneau and Lample 2019; Liu et al. 2020), and with backtranslation (BT; Sennrich et al., 2016b). However the literature is contradictory about the relative effectiveness of these techniques.

In this work, we note that previous work has differed in their training and test setups, and some research has used relatively small-sized models with only a few languages. In carefully controlled experiments, we use an MMT dataset with 100 translation directions, and 4 different test sets covering different domains. In addition, we experiment with model scale, ranging from 90M to 1.6B parameters. We find that monolingual data generally helps MMT, but models are surprisingly brittle to domain mismatches, especially at smaller model scales. BT is beneficial when the parallel, monolingual, and test data sources are similar but can be detrimental otherwise, while DAE is less effective than previously reported. Examining the effect of scale, we find it is important for both methods, particularly DAE. As scale increases, DAE transitions from underperforming the parallel-only baseline at 90M to converging with BT performance at 1.6B, and even surpassing it in low-resource.

A full description of this work can be found in our paper (Baziotis et al., 2023)

2.2.2 Question Translation Training for Better Multilingual Reasoning

Large language models have recently shown a strong ability to reason in English, but performance in other languages, especially more distant languages, still trails far behind (Shi et al., 2022). It is unsurprising, considering that their training data is predominantly composed of English text and instructions (Blevins and Zettlemoyer, 2022). To elicit LLM's multilingual performance, the previous approach typically follows the translate-training paradigm (Chen et al., 2023), which first translates English instruction data into non-English with a translation engine and then uses the multilingual data for instruction-tuning.

However, the translate-training has the following drawbacks: (1) translating English training data to numerous non-English languages incurs significant translation cost, especially considering the constant addition of large and complex instruction tuning sets (Yuan et al., 2023). (2) Additionally, it is hard for the translation engine to accurately translate lengthy, logical texts containing mathematical symbols in chain-of-thought responses, which can compromise the quality of translated data. Consequently, we explore the following research question in this paper: *Can we unlock the LLM's multilingual reasoning ability by teaching it to translate reasoning questions into English?*

In this paper, we focus on the multilingual mathematical reasoning task and explore the benefits of question alignment (QAlign), where we fine-tune the pre-trained LLM to translate reasoning questions into English with X-English parallel question data. This targeted, in-domain language alignment enables the subsequent effective utilization of English instruction data to unlock LLMs' multilingual reasoning abilities. Following question alignment, we implement response alignment by further fine-tuning the language-aligned LLM with cutting-edge English instruction data. Even though we use English-only supervised data, our alignment-enhanced LLM can achieve superior performance on non-English tasks with its transferable English expertise. Our method is illustrated in Figure 3.

To demonstrate the advantages of question alignment, we conduct experiments on challenging multilingual mathematical reasoning benchmarks, MGSM (Shi et al., 2022) and MSVAMP (Chen et al., 2023). We use two of the most advanced open-source LLMs, LLaMA2-7B and LLaMA2-

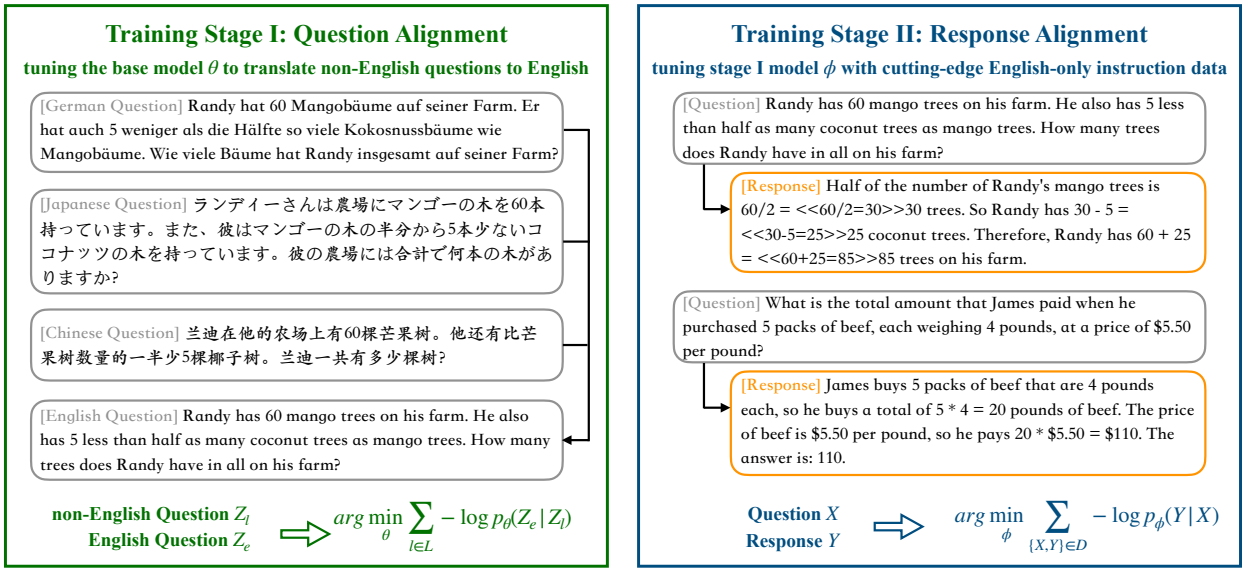

Figure 3: Illustration of our devised two-step training framework. At training stage I, we use a set of multilingual questions for translation training. At training stage II, we use cutting-edge English-only supervised data for fine-tuning. Due to the established language alignment in stage I, the LLM's proficiency in English can be transferred to non-English tasks.

13B (Touvron et al., 2023), as base models. Experiment results show that the inclusion of the question alignment stage brings an average improvement of up to 13.2% in multilingual performance. The performance improvement on low-recourse languages, e.g. Thai and Swahili, can be 30%-40%. Compared to the translate-training baseline, MathOctopus (Chen et al., 2023), which tuned with a multilingual version of GSM8K dataset, our alignment-enhanced LLMs achieves average performance improvement of 9.6% (7B) and 11.3% (13B) on MGSM. On the out-of-domain test set MSVAMP, our fine-tuned LLMs achieve 13.1% (7B) and 16.1% (13B) average accuracy improvement, also demonstrating our approach is robust to domain shift. In general, we observe that incorporating translated instruction data does benefit multilingual performance, but our question alignment strategy provides a more efficient and effective choice. In our analysis, we also present the effects of other implementations for performing language alignment and illustrate the importance of choosing the appropriate translation direction and domain during this phase of training.

The main contributions of this paper can be summarized as:

- We present a novel X-English question alignment finetuning step which performs targeted language alignment for the best use of the LLMs English reasoning abilities.
- We fine-tune open-source LLMs, LLaMA2-7B/13B, into strong multilingual reasoners, which beat the translate-training baseline by 9.6% (7B) and 11.3% (13B) on MGSM, by 13.1% (7B) and 16.1% (13B) on MSVAMP.
- We explore language alignment with other language directions (English-X), and types and domains of data, e.g. CoT responses and FLORES, and confirm our intuition that in fact X-English questions perform best.

A full description of this work can be found in our paper (Zhu et al., 2024).

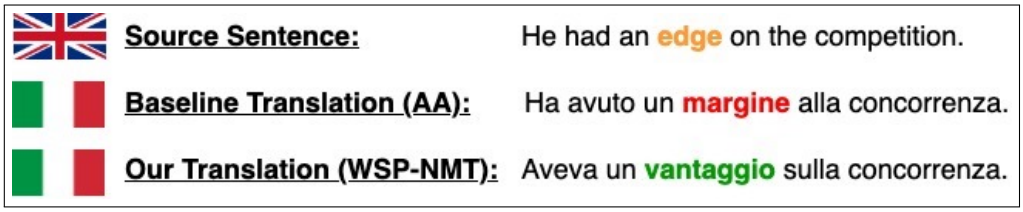

Figure 4: Italian translations of a sentence from the DiBiMT disambiguation benchmark (Campolungo et al., 2022) by: a) our main baseline, Aligned Augmentation (AA, Pan et al., 2021), and b) our approach, WSP-NMT. AA mistranslates the ambiguous word `edge` as `margine` *(border, rim)*. Due to 'sense-pivoted pretraining', WSP-NMT correctly translates it as `vantaggio` *(advantage)*.

2.2.3 Code-Switching with Word Senses for Pretraining in Neural Machine Translation

Lexical ambiguity is a long-standing challenge in Machine Translation due to polysemy being one of the most commonly occurring phenomena in natural language. Indeed, thanks to a plethora of context-dependent ambiguities (e.g. the word *run* could mean *run a marathon*, *run a mill*, *run for elections* etc.), words can convey very distant meanings, which may be translated with entirely different words in the target language. To deal with this challenge, traditional Statistical Machine Translation approaches tried to incorporate Word Sense Disambiguation (WSD) systems in MT with mostly positive results (Carpuat and Wu, 2007). These were followed by similar efforts to plug sense information in NMT frameworks (Liu et al., 2018). But, since the introduction of the Transformer, the task of disambiguation has largely been left to the attention mechanism.

In the last three years, though, many works have challenged the ability of modern-day NMT systems to accurately translate highly polysemous and/or rare word senses (Emelin et al., 2020). A likely explanation is that these models capture inherent data biases during pretraining. This particularly holds for the pretraining paradigm of denoising code-switched text[4] — most notably, Aligned Augmentation (AA, Pan et al., 2021), where, during the pretraining phase, input sentences are noised by substituting words with their translations from multilingual lexicons, and NMT models are then tasked to reconstruct (or 'denoise') these sentences. AA and subsequent works (Iyer et al., 2023c) show the benefits of code-switched pretraining for high- and low-resource, supervised and unsupervised translation tasks. Despite their success, a major limitation of these substitution mechanisms is that they are unable to handle lexical ambiguity adequately, given their usage of *'sense-agnostic'* translation lexicons. In fact, in most of these works, substitutions for polysemes are chosen randomly, regardless of context (Pan et al., 2021).

In an effort to introduce knowledge grounding at the word sense level during pretraining and potentially minimise data errors, enhance convergence, and improve performance, we propose the notion of ***'sense-pivoted pretraining'*** – to move code-switched pretraining from the *word level* to the *sense level*. Specifically, we propose an approach called Word Sense Pretraining for Neural Machine Translation (WSP-NMT) which first disambiguates word senses in the input sentence, and then code-switches with sense translations for denoising-based pretraining. Figure 4 provides an intuition of how integrating disambiguation in pretraining helps our model handle ambiguous words better, avoiding defaulting to more frequent senses, and reducing errors in translation.

Indeed, our experiments on using WSP-NMT yield significant gains in multilingual NMT – about +1.2 spBLEU and +0.02 COMET22 points over comparable AA baselines in high-resourced setups. Among other interesting performance trends, we observe that our margin of improve-

[4] In this work, we refer to this family of approaches as 'code-switched pretraining' for brevity

ment increases substantially as we move towards low-resource (+3 to +4 spBLEU) and medium-resource (+5 spBLEU) settings. Lastly, for more fine-grained evaluation, we also compare our models on the DiBiMT disambiguation benchmark (Campolungo et al., 2022) for Italian and Spanish, and note accuracy improvements of up to 15% in the challenging task of verb disambiguation.

Our key novel contributions are, thus, as follows:

1. We show how incorporating WSD in NMT pretraining can outperform the widely used paradigm of lexicon-based code-switching.
2. We demonstrate how reliable structured knowledge can be incorporated into the multilingual pretraining of NMT models, leading to error reduction and improved performance.
3. We evaluate the robustness of WSP-NMT to scale to various challenging data and resource-constrained scenarios in NMT and point out its efficacy in low-resource and zero-shot translation tasks.
4. Finally, we evaluate the disambiguation capabilities of our models on the DiBiMT benchmark and contribute a fine-grained understanding of the scenarios where WSP-NMT helps resolve lexical ambiguity in translation.

A full description of this work can be found in our paper (Iyer et al., 2023a).

2.2.4 Is Modularity Transferable? A Case Study through the Lens of Knowledge Distillation

The rise of Modular Deep Learning (Pfeiffer et al., 2023) showcases its potential in various Natural Language Processing applications. The use cases start with parameter-efficient fine-tuning (PEFT) methods through multi-task transfer learning up to the cross-lingual and cross-task adaptation of Pre-trained Language Models (PLMs). However, the current modularity is attached to its base model, that is, model-specific modularity. We ask whether current modular approaches are transferable between models and whether we can transfer the modules from more robust and larger PLMs to smaller ones. In this work, we aim to fill this gap via a lens of Knowledge Distillation (Hinton et al., 2015), showing an extremely straightforward approach to transferring pre-trained, task-specific PEFT modules between same-family PLMs. Moreover, we propose a method that allows the transfer of modules between incompatible PLMs without any change in the inference complexity. The experiments on Named Entity Recognition, Natural Language Inference, and Paraphrase Identification tasks over multiple languages and PEFT methods showcase the initial potential of transferable modularity.

A full description of this work can be found in our paper (Klimaszewski et al., 2024).

2.3 Retrieval augmented generation

While effective in many contexts, traditional generation methods often struggle with generating content that requires factual accuracy or specific knowledge that goes beyond their training data. Retrieval augmented generation(RAG) techniques, focus on dynamically incorporating external information during generation in order to overcome these challenges. In the following two works, we propose retrieval-augmented methods and demonstrate how they can be integrated with large language models to demonstrably improve generation performance across tasks and metrics.

2.3.1 Retrieval-augmented multilingual knowledge editing

Large Language Models (LLMs) are being used as sources of factual knowledge for search engines and other downstream tasks. Despite their considerable progress, knowledge generated by LLMs can be incorrect or become obsolete in a changing world. Pre-training from scratch or fine-tuning LLMs to adapt them to new knowledge is computationally expensive and not guaranteed to work. Knowledge editing (KE) methods (Zhu et al., 2020; Cao et al., 2021) have been proposed as effective and economic alternatives to fine-tuning when specific factual knowledge needs to be added or updated. KE involves either updating the parameters of a model (Dai et al., 2022a; Mitchell et al., 2022a; Meng et al., 2022, 2023; Dai et al., 2022b) or adding extra components to an LLM (Mitchell et al., 2022b; Zheng et al., 2023). For example, KE can be used to correct the answer to this question *"Who is the foreign secretary of the UK?"* from *"James Cleverly"* (true until mid November 2023) to *"David Cameron"*, who has recently been appointed to the post.

Despite significant interest in this problem, current research on KE predominantly concentrates on a monolingual setting, where both the injected knowledge and the subsequent queries to the LLM are in English (Mitchell et al., 2022a; Meng et al., 2022, 2023; Mitchell et al., 2022b; Zheng et al., 2023). Companies serving a multilingual customer base need to consider the multilingual KE case, where KE is done in one language and this propagates to queries and answers in all other languages. While Wang et al. (2023a) explored the cross-lingual applicability of knowledge editing to the English-Chinese cross-lingual scenario, their primary focus was to highlight the challenges rather than develop a functional KE approach in a multilingual setting.

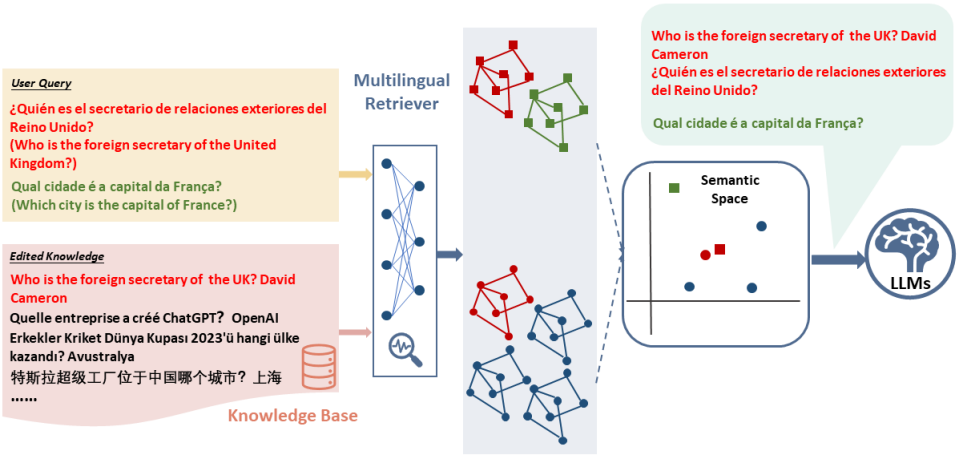

Figure 5: ReMaKE attaches in-context knowledge to an LLM prompt when it is retrieved (red example where the edited knowledge is in English and a user query is in Spanish) from a customer-defined multilingual knowledge base. When no edited knowledge is retrieved (green example) the prompt is passed to the LLM unchanged.

Drawing inspiration from in-context learning (ICL), in-context knowledge editing (IKE) uses prompts to edit factual knowledge. It is noted that IKE is so far the only method demonstrating positive results in the cross-lingual KE task setting (Wang et al., 2023a). However, IKE requires explicit provision of new knowledge every time an LLM is used, confining its practicality and scalability in real-world applications. In addition, IKE suffers when irrelevant facts are included in the prompt (Wang et al., 2023c) especially in scenarios where a substantial number of facts are being edited.

In this paper, we propose **Retrieval-Augmented Multilingual Knowledge Editor (ReMaKE)** that integrates multilingual retrieval from a knowledge base with in-context learning. ReMaKE concatenates the retrieved knowledge from an external database with a user query to create the prompt. The proposed multilingual retriever grounds the ReMaKE to the retrieved accurate and up-to-date information highly relevant to user queries, therefore effectively mitigating the contextual interference due to irrelevant context. In this way, the generated prompts are able to guide the LLMs in producing accurate responses associated with the injected knowledge. ReMaKE leverages a knowledge base's ability to scale to further enhance IKE's knowledge editing performance in real-world application scenarios where large volumes of edits are in scope. Figure 5 shows the architecture of the proposed retrieval-augmented multilingual knowledge editor. Our main contributions are listed below:

- **Multilingual knowledge editing**: ReMaKE extends the scope of knowledge editing practices across language boundaries. Given that the multilingual knowledge base and multilingual retriever operate independently to a specific LLM, ReMaKE is a **plug-and-play** tool applicable to any LLM. It is **scalable**, capable of editing a large number of knowledge. Experiments show ReMaKE outperforms baseline methods by a significant margin in the average accuracy score (up to +40.53%).
- **Multilingual editing dataset**: We build a machine-translated multilingual knowledge editing dataset (**MzsRE**) in 12 languages: English, Czech, German, Dutch, Spanish, French, Portuguese, Russian, Thai, Turkish, Vietnamese, and Chinese using the zsRE testset (Levy et al., 2017). The dataset will be made available to the community.

A full description of this work can be found in our paper (Wang et al., 2023b).

2.3.2 Empirical Assessment of kNN-MT for Real-World Translation Scenarios

The remarkable advances in neural models have brought significant progress in the field of machine translation (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). However, current systems rely heavily on a fully-parametric approach, where the entire training data is compressed into the model parameters. This can lead to inadequate translations when encountering rare words or sentences outside of the initial training domain (Koehn and Knowles, 2017), requiring several stages of fine-tuning to adapt to data drift or to new domains.

By combining the advantages of parametric models with non-parametric databases built from parallel sentences, retrieval-augmented models showed to be a promising solution, particularly in domain adaptation scenarios (Gu et al., 2018; Zhang et al., 2018; Bapna and Firat, 2019; Meng et al., 2021; Zheng et al., 2021; Jiang et al., 2021; Martins et al., 2022a,b).

One notable example is the *k*-Nearest Neighbor Machine Translation model (*k*NN-MT) (Khandelwal et al., 2021), known for its simplicity and very promising results. The model first creates a token-level datastore using parallel sentences, and then it retrieves similar examples from the database during inference, enhancing the generation process via interpolation of probability distributions.

However, despite its potential, the *k*NN-MT model has yet to be tested in real-world scenarios. Previous studies have primarily focused on evaluating it using only the BLEU metric, which correlates poorly with human judgments. In order to gain a deeper understanding of when and how

*k*NN-MT can be effective, we conduct a thorough analysis on various datasets which comprise 4 different language pairs and 3 different domains, using BLEU (Papineni et al., 2002; Post, 2018), COMET (Rei et al., 2020), and Multidimensional Quality Metrics (MQM) – quality assessments obtained from the identification of error spans in translation outputs by experts (Lommel et al., 2014; Freitag et al., 2021). Additionally, we assess the impact of datastore size on translation quality and determine the necessary number of entries for configuring the datastore index.

	En-De (2)				En-Fr				En-Ko			
	MINOR	MAJOR	CRITICAL	MQM	MINOR	MAJOR	CRITICAL	MQM	MINOR	MAJOR	CRITICAL	MQM
Base Model	1301	896	439	61.24	499	237	266	88.42	713	185	28	75.23
*k*NN-MT	928	417	75	86.22	**335**	**116**	137	93.77	527	**95**	6	85.72
Fine-tuned	982	471	72	85.03	377	131	**3**	**97.14**	513	101	**3**	85.56
Fine-tuned + *k*NN-MT	**800**	**391**	**62**	**88.03**	363	118	5	96.87	**466**	99	5	**85.97**

Table 3: Error severity counts and MQM scores.

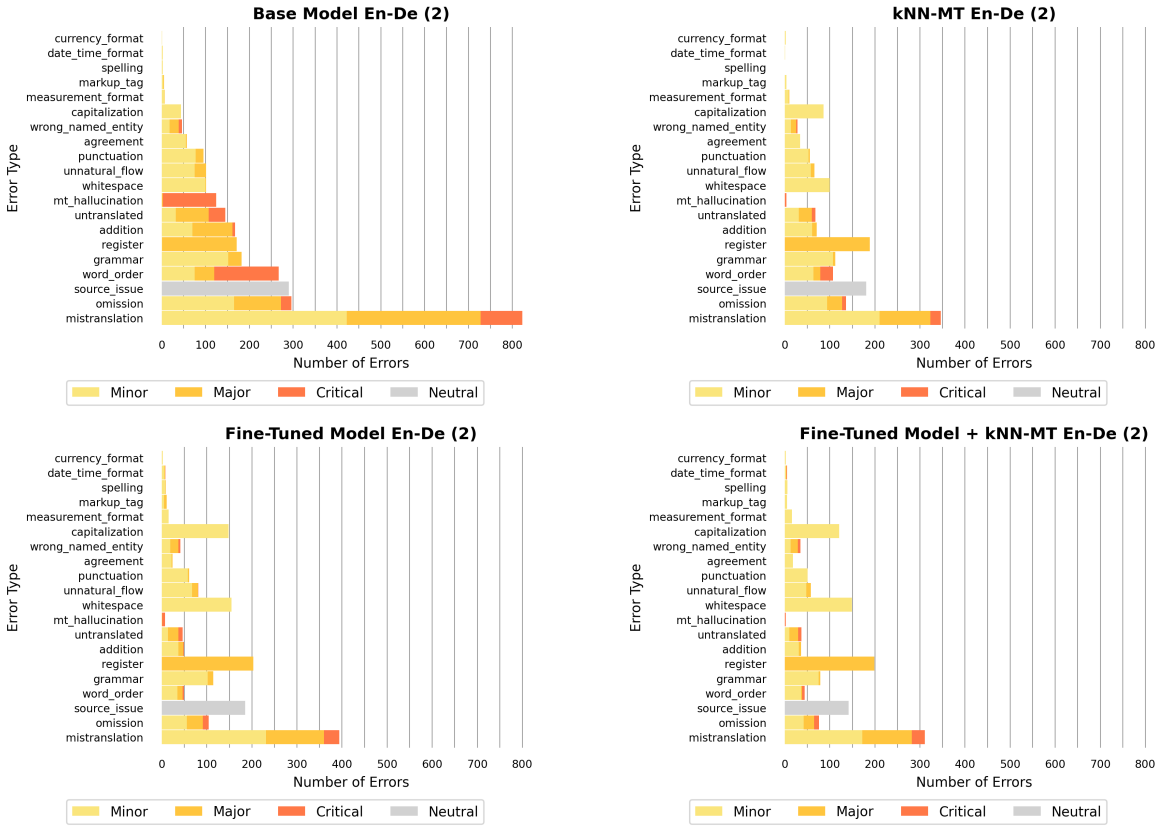

Figure 6: Error typology and severity level breakdown for the En-De (2) test set.

Our results demonstrate that kNN-MT significantly enhances the pretrained model's performance across all tested datasets, showcasing increased BLEU and COMET scores. While kNN-MT did not outperform full or adapter-based fine-tuning according to automatic metrics, the MQM evaluations, shown in Table 3 and Figure 6, provide a different perspective, indicating that kNN-MT slightly surpasses fine-tuning in two out of three datasets. Moreover, our findings reveal that larger datastores improve translation quality and that even with a limited amount of data, it is possible to create an effective datastore and index for kNN-MT.

A full description of this work can be found in our paper (Martins et al., 2023).

2.4 Analysis and improvement of neural translation capabilities

The works in this section focus on specific intricacies of machine translation that can affect performance and robustness. Namely, §2.4.1 employs two cheating methodologies to detect and analyse the characteristics of hard problems. §2.4.2 on the other hand targets semantic ambiguity mitigation with LLMs and proposes methods and datasets to advance this topic further.

2.4.1 Cheating to Identify Hard Problems for Neural Machine Translation

We identify hard problems for neural machine translation models by analyzing progressively higher-scoring translations generated by letting models cheat to various degrees. If a system cheats and still gets something wrong, that suggests it is a hard problem. We experiment with two forms of cheating: providing the model with a compressed representation of the target as an additional input, and fine-tuning on the test set. Contrary to popular belief, we find that the most frequent tokens are not necessarily the most accurately translated due to these often being function words and punctuation that can be used more flexibly in translation, or content words which can easily be paraphrased. We systematically analyze system outputs to identify categories of tokens which are particularly hard for the model to translate and find that this includes certain types of named entities, subordinating conjunctions, and unknown and foreign words. We also encounter a phenomenon where words, often names, which were not infrequent in the training data are still repeatedly mistranslated by the models — we dub this the Fleetwood Mac problem.

A full description of this work can be found in our paper (Pal and Heafield, 2023).

2.4.2 Towards Effective Disambiguation for Machine Translation with Large Language Models

Resolving semantic ambiguity has long been recognised as a central challenge in the field of machine translation. Recent work on benchmarking translation performance on ambiguous sentences has exposed the limitations of conventional Neural Machine Translation (NMT) systems, which fail to handle many of these cases. Large language models (LLMs) have emerged as a promising alternative, demonstrating comparable performance to traditional NMT models while introducing new paradigms for controlling the target outputs. In this paper, we study the capabilities of LLMs to translate ambiguous sentences containing polysemous words and rare word senses. We also propose two ways to improve the handling of such ambiguity through in-context learning and fine-tuning on carefully curated ambiguous datasets. Experiments show that our methods can match or outperform state-of-the-art systems such as DeepL and NLLB in four out of five language directions. Our research provides valuable insights into effectively adapting LLMs for disambiguation during machine translation.

A full description of this work can be found in our paper (Iyer et al., 2023b).

2.5 Quality Evaluation for Machine Translation

In this section, we report works that focus on evaluating the performance of models tuned for machine translation. We report works that adapt neural network architectures or LLMs to predict MT quality and detect errors at different levels of granularity. Additionally, we describe two core shared tasks in this domain (*WMT Metrics* [5] and *WMT Quality Estimation* [6] shared tasks) for which UTTER partners are actively involved in the co-organisation.

2.5.1 The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation

In this work, we explore the advancement of automatic evaluation in machine translation (MT) focusing on the potential for fine-grained error annotations. We propose AUTOMQM, a novel technique that uses the reasoning capabilities of large language models (LLMs) to identify and categorize errors in translations. We thus address the limitations of traditional metrics which have focused on producing a single quality score, lacking the depth provided by more detailed schemes like the Multidimensional Quality Metrics (MQM) that allow humans to annotate individual errors. AUTOMQM fills this gap by prompting LLMs, such as PaLM and PaLM-2, to not only predict scores but to also pinpoint and classify translation errors, thus offering both a quantitative evaluation and qualitative insights.

Our approach evaluates the potential of LLMs in score prediction through prompting and examines the impact of labelled data via in-context learning and fine-tuning. AUTOMQM's introduction allows us to leverage the MQM framework to produce scores based on identified errors, providing a level of interpretability previously unavailable. This approach results in significant performance improvements, especially with larger models, and offers interpretability by aligning error spans with human annotations.

Our main contributions are summarised as:

- AUTOMQM enables LLMs to effectively identify and categorize translation errors, offering a substantial improvement over mere score prediction methods.
- Fine-tuning LLMs with human judgment data significantly enhances their performance in evaluating translations, particularly at the segment level.
- AUTOMQM allows LLMs to generate rich, MQM-like annotations, outperforming score prediction approaches in segment-level evaluation.
- Annotations predicted by LLMs through AUTOMQM accurately identify a considerable portion of words involved in major errors, providing valuable feedback for MT system improvements.

Our work demonstrates the value of employing LLMs not just for producing quality scores but for providing detailed feedback on translations, thereby advancing the evaluation of MT systems. By integrating AUTOMQM into the evaluation process, we pave the way for a deeper understanding and enhancement of machine-generated translations, highlighting the importance of fine-grained analysis in the progress of MT evaluation.

[5] https://www2.statmt.org/wmt24/metrics-task.html
[6] https://www2.statmt.org/wmt24/qe-task.html

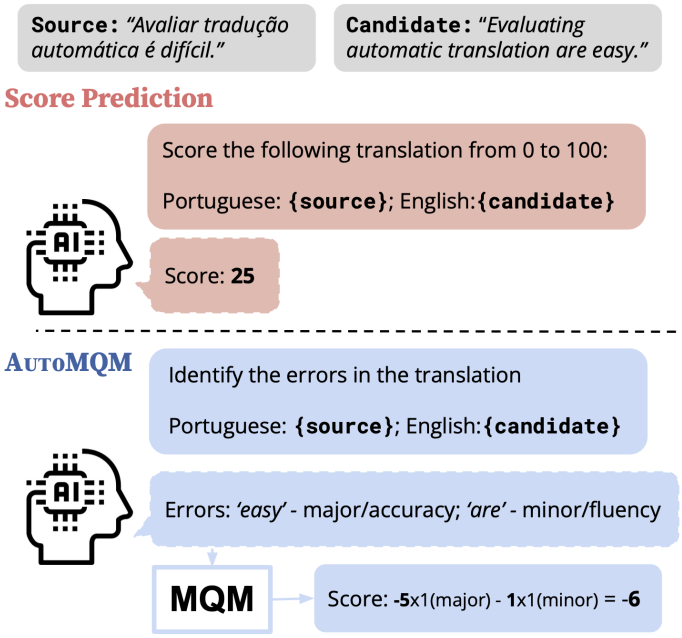

Figure 7: Illustration of the AUTOMQM process, showcasing how LLMs are prompted to assess the quality of a translation by identifying and classifying errors, and using the MQM framework to produce a score.

More details about our work can be found in our EMNLP 2023 paper (Fernandes et al., 2022).

2.6 CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task

In our efforts for the WMT 2022 Shared Task on Quality Estimation, we introduced CometKiwi, a novel framework that combines the predictive capabilities of COMET with the estimator architecture of OpenKiwi. This fusion aims to enhance quality estimation by integrating word-level tagging and explanation extraction, informed by pretraining on Direct Assessments and fine-tuning with both sentence-level and word-level shared task data.

Our approach begins with pretraining on a large corpus assembled from previous WMT-Metrics shared tasks, followed by fine-tuning on the WMT-QE 2022 data. This process is designed to equip CometKiwi with the ability to generalize across multiple languages and adapt to new ones through few-shot training.

The overall architecture of our models is shown in Figure 8. The machine translated sentence $\boldsymbol{t} = \langle t_1, ..., t_n \rangle$ and its source sentence counterpart $\boldsymbol{s} = \langle s_1, ..., s_m \rangle$ are concatenated and passed as input to the encoder, which produces d-dimensional hidden state vectors $\boldsymbol{H}_0, ..., \boldsymbol{H}_L$ for each layer $0 \leq \ell \leq L$, where $\boldsymbol{H}_i \in \mathbb{R}^{(n+m)\times d}$, where $\ell = 0$ corresponds to the embedding layer. Next, all hidden states are fed to a scalar mix module (Peters et al., 2018) that learns a weighted sum of the hidden states of each layer of the encoder, producing a new sequence of aggregated hidden states, $\boldsymbol{H}_{\text{mix}}$.

For sentence-level models, the hidden state of the first token (<cls>) is used as sentence representation $\boldsymbol{H}_{\text{mix},0} \in \mathbb{R}^d$, which, in turn, is passed to a 2-layered feed-forward module in order to get a sentence score prediction $\hat{y} \in \mathbb{R}$. For word-level models, we first retrieve the hidden state vectors associated with the first word piece of each machine-translated token and then pass them to a linear projection to get word-level predictions $\hat{y}_i \in \{\text{OK}, \text{BAD}\}, \forall_{1 \leq i \leq n}$. Moreover, attention matrices

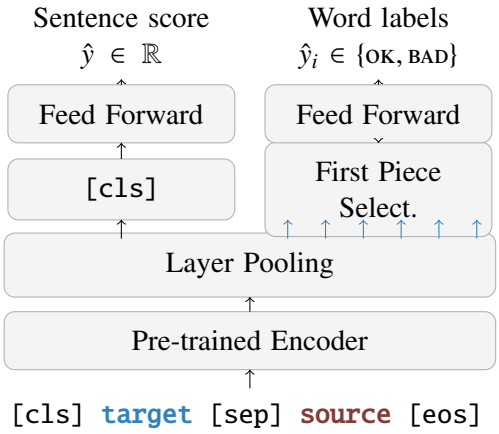

Figure 8: General architecture of CometKiwi for sentence-level (left part) and word-level QE (right part).

$\boldsymbol{A}_{1,1}, ..., \boldsymbol{A}_{L,H}$ for all layers and heads are also recovered as a by-product of the forward propagation.

Our key contributions are as follows:

- The creation of CometKiwi, which marries the strengths of COMET's training features with OpenKiwi's architecture, adding word-level sequence tagging for enhanced QE.
- Validation of pretraining importance on annotations from metrics shared tasks for improving downstream task performance.
- Improvement in quality estimation for language pairs new to the training data, demonstrating the effectiveness of our methodology in multilingual settings, even in zero-shot conditions.
- The proposal of a novel interpretability method that leverages attention and gradient information for better explanation extraction.

A full description of this work can be found in our paper (Rei et al., 2022b)

2.6.1 COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task.

This work presents collaborative participation in the WMT 2022 Metrics Shared Task. We introduce COMET-22, an ensemble approach that incorporates both sentence-level scoring and fine-grained error detection through word-level tagging, leveraging advancements in COMET estimator models and Multidimensional Quality Metrics (MQM).

Our research advances metric development for machine translation evaluation by:

- Proposing a multitask model architecture that enhances the utilization of MQM data for both sentence-level scoring and word-level tagging, applicable with and without reference translations.
- Demonstrating the effectiveness of model ensembling from varied annotations in achieving higher correlations with human judgments and improved robustness against translation errors.

- Validating the increasing competitiveness of reference-free evaluation methods against traditional reference-based evaluations.

Our findings underscore significant improvements over previous state-of-the-art metrics in terms of correlation with MQM annotations and the ability to detect critical errors. Notably, our approach shows enhanced performance in identifying and penalizing deviations in named entities, numbers, and overall meaning.

COMET-22 represents a significant step forward in the development of metrics for machine translation evaluation. By effectively integrating sentence-level scores with word-level tags and utilizing MQM data, our submissions not only show improved correlations with human judgments but also a heightened sensitivity to critical translation errors. Our work paves the way for further innovations in the field of machine translation metrics.

A full description of this work can be found in our paper (Rei et al., 2022a).

2.6.2 Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task.

This work extends the CometKiwi model, employing a large-scale dataset to fine-tune our models for enhanced performance across sentence and word-level quality prediction, as well as fine-grained error-span detection tasks.

We use the 2017 to 2019 data from the WMT shared tasks, annotated with direct assessments, to construct generic models for quality estimation, which are then fine-tuned for specific tasks. After having obtained the generic models, we will train models for each separate stream of the shared task, i.e., sentence-level, word-level or error span prediction. To do so, we consider the multi-task optimization from wherein sentence scores can be used alongside supervision from word-level tags.

While we also use InfoXLM L [7] as in Rei et al. (2022a), we scale up our multilingual encoders, using also XLM-R XL [8], and XLM-R XXL [9]. As the scaling-up resulted in improved performance across tasks, we publicly release two of our best models for research purposes (COMETKIWI-XL [10], and-XXL[11]). To the best of our knowledge, these are the largest open-source QE models publicly released.

Our overall contributions are manifold:

1. We introduced a series of novel approaches for multilingual quality estimation, demonstrating the top-ranked performance across all tasks.

2. We explore different approaches to predict the span of erroneous translations along with their error severities (three-way classification: OK, MINOR, MAJOR);

3. We have made two of our large, top-performing models publicly available, encouraging further research and development in the field.

[7] https://huggingface.co/microsoft/infoxlm-large
[8] https://huggingface.co/facebook/xlm-roberta-xl
[9] https://huggingface.co/facebook/xlm-roberta-xxl
[10] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl
[11] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

A full description of this work can be found in our paper (Rei et al., 2023).

2.6.3 Quality Estimation Shared Tasks

A significant contribution to the advancement of the field in the evaluation of the machine translation models comes via the organisation of the WMT Quality Estimation shared tasks, which contribute significant annotated data, and motivate the development of novel models and methodologies. We have been co-organising the tasks, both contributing data annotations (Unbabel) and being involved in the task design, planning and implementation.

In recent editions, we have been continuously introducing new language pairs, emphasising both UTTER languages (German, Portuguese, and this year, Spanish) as well as lower resource languages that test the robustness of submitted models. Additionally, we incorporated quality annotations with Multidimensional Quality Metrics (MQM) for sentence- and word-level quality scores, lately targeting more fine-grained quality estimation. We are also encouraging multilingual and zero-shot approaches, to further explore the capabilities of LLMs in the field.

For more information, we direct the readers to the WMT Quality Estimation Findings papers for 2022 and 2023 respectively (Zerva et al. (2022); Blain et al. (2023)).

2.6.4 Metrics Shared Tasks

Similar to the quality estimation shared tasks reported in the previous section, the Metrics shared task covers a key aspect of MT evaluation, focusing not only on segment-level evaluation but also on comparing and evaluating MT models on the system level. The task invites participants to submit neural or traditional metrics that evaluate translation instances and systems in multilingual, multi-domain settings. Employing a continuously updated, multi-dimensional analysis of the results we are able to compare and evaluate available metrics and draw conclusions on their performance (e.g. emphasising the performance superiority of neural metrics in recent years). Besides the main evaluation, submitted metrics are stress-tested against a set of challenge sets (also submitted by participants) to assess their robustness. In the latest edition, we have seen the proposal of new evaluation metrics and models that better align with human judgment, reflecting a deeper understanding of language intricacies. These advancements underscore the community's ongoing commitment to enhancing the accuracy, reliability, and interpretability of MT systems.

For more information, we direct the readers to the WMT Metrics Findings papers for 2022 and 2023 respectively (Freitag et al. (2022b, 2023)).

Plans for future work

This task holds increased importance for UTTER's goals and we expect to continue working on it in the second half of the project as originally planned. We will specifically focus more on adaptation for large language models and instruction-tuning as well as alignment methods for multilingual tasks.

3 Task T4.2: Contextualisation and emotion tracking (IT*, UEDIN, UVA, UNB)

Proposal highlights

The goal of this task is to develop language systems for conversational data that exploit the context of the conversation, such as the previous utterances (dialogue context), meta-information about the speaker, and the emotional state of the speaker. The key highlights from the proposal are listed below.

- Proposal of refined archetypal emotions for conversational emotion recognition that considers both speech and text signals, as well as conversation history.
- Investigation of new methods for flexible and efficient ways to encode context.
- New evaluation methods for context usage and impact on translation quality.

Summary of completed work

This is a key task for the conversational agents developed within UTTER. This first half of the proposal has seen advancements across all three key areas described above, with increased emphasis on analysing the impact of context usage on translation quality, and the development of novel context-aware methods. The latter dimension is increasingly important with the development of large language models that can receive extended text as input, for which however we still have limited knowledge regarding its impact on generation quality and consistency.

3.1 Context for speech-to-text translation

In this section, we report work that explores whether direct speech-to-text models can benefit from additional context contained in audio signals, such as prosody, compared to using cascade approaches.

3.1.1 Prosody in Cascade and Direct Speech-to-Text Translation: a case study on Korean Wh-Phrases

In speech translation (ST), a *direct* model translates from the source audio to the target text without using an intermediate discrete representation. This is in contrast to a *cascade* system which typically involves coupling an ASR (automatic speech recognition) model with a text-to-text MT (machine translation) model. One of the advantages claimed for direct models is that the translation model has access to the full audio signal, as opposed to the direct system where the audio is discarded before translating. In this work, we wanted to test whether direct models could actually exploit this access to the audio signal.

Our study focused on Korean *wh*-phrases, for which prosodic features are necessary to produce the correct translation. In these constructions, written Korean may not distinguish between statements, yes/no questions, *wh*-questions or other possibilities. Using a test set containing ambiguous Korean *wh*-phrases, we compared the performance of strong direct and cascade ST systems using

contrastive evaluation on the English references. We were able to show that the direct system was indeed able to exploit the prosody of the source, showing a 12.9% improvement in overall accuracy in the ambiguous cases. However, the accuracy still remains low in absolute terms, showing that this is a difficult problem for ST systems. To our knowledge, this is the first study to provide quantitative evidence that direct ST models can effectively leverage prosody.

A full description of this work can be found in our paper (Zhou et al., 2024b).

3.2 Context-aware machine translation

This strand of work focuses on improving MT with respect to the better use of context around a source sentence that is to be translated, thus aiming to obtain improved translations that better fit the domain and intended meaning in the original text. This is particularly relevant for the translation of long documents and dialogue data. In particular, within the context of UTTER, this strand of work relates to the improvement of models used for both of our intended use cases, namely the customer-support assistant (WP 7.1) and the meeting assistant (WP 7.2), since they both involve potentially long conversation and dependencies. Hence work in this section focuses on (a) the evaluation of existing models with respect to the use of context and robustness to different discourse phenomena, (b) the improvement of annotation and evaluation methodology that could better incorporate context and (c) the improvement of MT models so that they can better account for context during translation.

3.2.1 When Does Translation Require Context?

In this work, we expanded previous findings in the area of context-aware machine translation (MT) Fernandes et al. (2021), focusing on a comprehensive analysis of discourse, driven by the hypothesis that a deeper understanding and integration of discourse phenomena can markedly improve the quality of translations.

We propose the Multilingual Discourse-Aware (MUDA) benchmark, a framework equipped with taggers to identify and evaluate model performance on various discourse phenomena across datasets. Our approach leverages the Pointwise Cross-Mutual Information (P-CXMI) metric, designed to pinpoint translations that benefit from contextual information. This metric enables us to systematically explore and validate the complexity of well-known phenomena while also identifying new challenges, such as the consistency of verb forms.

Our main contributions are summarised below:

- We introduce the Multilingual Discourse-Aware (MUDA) benchmark, an innovative framework designed to identify and evaluate the performance of MT models on various discourse phenomena across datasets. We present the distribution of annotated MuDA tags on TED test data in Figure 9.
- The P-CXMI metric is proposed to systematically identify instances where translation benefits from contextual information, leading to a data-driven exploration of discourse phenomena across 14 language pairs (see also Figure 9).
- A thorough analysis highlights the limited improvements offered by current context-aware MT models over context-agnostic models, suggesting areas for future enhancement. Our

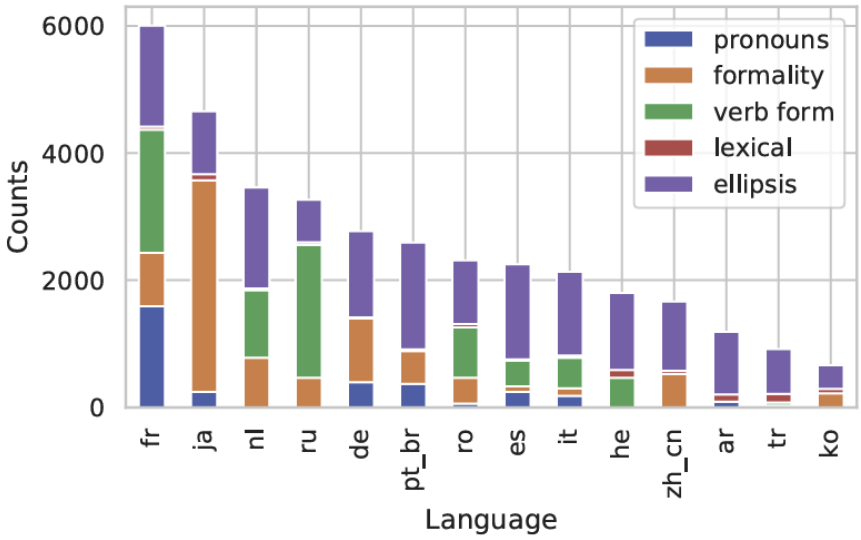

Figure 9: Distribution of MuDA tags per language pair on TED test data.

benchmarks demonstrate DeepL's superior performance in handling discourse phenomena compared to both its sentence-level ablation and Google Translate.

A full description of this work can be found in our ACL paper by Fernandes et al. (2023b), which was also awarded the **best resource paper award**.

3.2.2 Measuring Context Utilization in Document-Level MT Systems

In this work, we adopt an interpretable approach to context utilization evaluation. We evaluate models based on the ability to use the correct context, and not only the ability to generate a correct translation without necessarily utilizing the context.

Our contributions are the following:

- We perform a perturbation-based analysis on document-level models and find that single-encoder concatenation models are able to make use of the correct context vs. a random context.
- We propose the use of attribution scores of *supporting context* to evaluate correct context utilization. Analyzing the pronoun resolution phenomenon as a case study, we find that sentence-level models and single-encoder context-aware models are better than multi-encoder models in terms of the amount of attribution pronoun's antecedents have to generating the pronoun.
- We propose the use of automatically annotated *supporting context* as an alternative to human-annotated context for attribution evaluation. We show that, despite noise in automatic annotation, results are consistent with human-annotated context, paving the way towards efficient use of linguistic expertise in document-level translation evaluation.
- We highlight the importance of using a discourse rich dataset when evaluating the ability of models to handle context-dependent discourse phenomena.

A full description of this work can be found in our paper Mohammed and Niculae (2024).

3.2.3 Context-aware Neural Machine Translation for Dialogue

This work attempts to address the shortcomings of existing multilingual NMT models, with respect to taking into account context. It focuses on dialogue, and specifically business-related dialogue scenarios for English-Japanese translation. This setup allows us to explore specific discourse phenomena, including the use of formality across languages, as well as to experiment with extra-sentential context that describes the speaker or the conversation scene setup.

Our approach involves fine-tuning a pre-trained mBART model exploring different ways of encoding context. We experiment with source- and target-side context in the form of concatenated preceding sentences and propose a novel attention mechanism (CoAttMask) to better encode source-side context on the encoder side, without deteriorating the decoding quality. Furthermore, we experiment with encoding speaker-related and scene-related information as additional context to the source side, hypothesizing their potential to improve translation accuracy. See Figure 10 for an example of different context encodings.

We evaluate our context-aware models using BLEU and COMET scores for overall performance and employ Conditional Cross-Mutual Information (CXMI) for assessing context usage. Additionally, we use the extended version of p-CXMI Fernandes et al. (2023b) discussed in §3.2.1 to perform a focused analysis on the translation of Japanese honorifics, a critical aspect of En-Ja business dialogues. Our experiments reveal that:

- Models leveraging both preceding sentences and extra-sentential context (speaker turn and scene type) show improved translation quality. CXMI scores increase with context size, indicating more effective context utilization.
- The inclusion of speaker and scene information as additional context on the source side enhances model performance, particularly for larger context sizes.
- Our focused analysis on honorifics translation demonstrates that the provided context helps the model attribute higher probability to the correct honorific expressions, underscoring the importance of context in handling language-specific discourse phenomena.

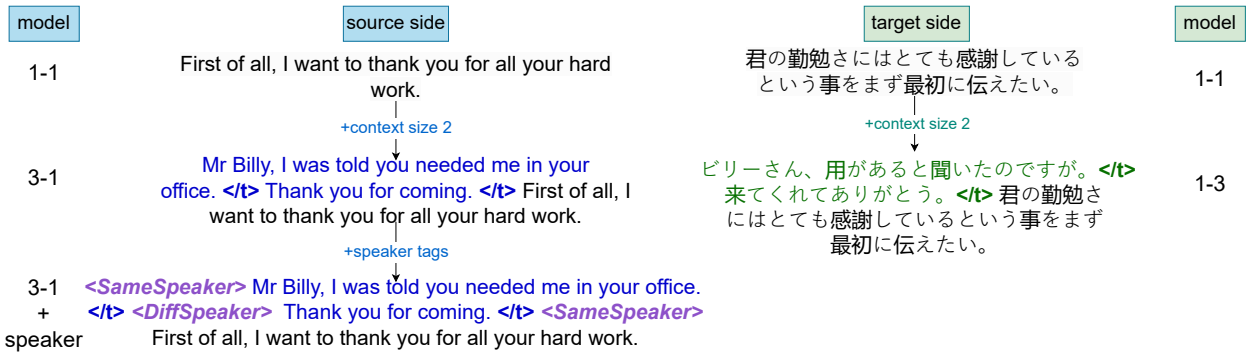

Figure 10: Context-extended inputs on source and target side. Coloured text corresponds to added context, **bold** signifies context separators and ***bold-italics*** speaker-related context tags.

Overall, this work underscores the significance of context in improving NMT for English-Japanese business dialogues, demonstrating that models can indeed leverage context to enhance translation quality.

A full description of this work can be found in our paper (Honda et al., 2023).

3.2.4 Context and emotion annotations in dialogue setups

In this section, we summarise works that focus on dialogue translation. Specifically, work in this first half of the project focused on redefining annotations for dialogue quality and emotion. The annotated data was also used for the chat translation shared task (to be repeated in 2024 for the UTTER languages [12]), which aims to motivate further advancements in the field of chat translation, as described in §3.2.7.

3.2.5 A Context-Aware Annotation Framework for Customer Support Live Chat Machine Translation

In this work we address the challenges and limitations of current MT quality assessment frameworks in capturing context-induced errors in customer support live chat scenarios. By revising the MQM framework to include contextual triggers, we aim to enhance quality evaluation for context-aware MT systems, underscoring the importance of context in achieving high-quality translations.

We used the MAIA corpus as prepared for the WMT 2022 Shared Task on Chat Translation Farinha et al. (2022), featuring authentic bilingual interactions from customer support chats. This dynamic and informal content, often filled with abbreviations, emoticons, idiomatic expressions, and errors, motivated our analysis. Our objectives were to comprehend how context is integrated within texts, identify lexical structures that carry contextual meanings, develop an annotation framework to assess context in documents effectively, and initiate the creation of a multilingual test suite. This suite aims to incorporate contextual annotations, addressing the complexities of real-world customer support data.

Our methodology involves the revision of the MQM framework Lommel et al. (2014) by incorporating nine new annotation categories that allow for the mapping of MT errors to specific contextual phenomena within source documents. This approach enables a more nuanced and comprehensive error identification process, highlighting the critical and major severity of many contextual errors.

We demonstrate the enhanced error identification capacity of the revised MQM framework, covering a significantly greater number of errors than traditional applications. We also highlight the limitations of existing evaluation metrics in detecting contextual errors, emphasizing the necessity of integrating our contextual annotation framework to better assess the impact of context on MT quality. Our framework shows significant gains of an average of 61% more contextual error coverage than more conventional QA metrics, highlighting the fact that most of such contextual errors are deemed as critical and major, thus strengthening our beliefs that the field of quality assessment for context-aware MT is far from being effectively dealt with, on the one hand, and that contextual error severely compromise MT outputs, on the other hand.

Our work contributes to the field by revising the MQM framework to better capture context-induced MT errors, providing a detailed analysis of these errors' severity, and highlighting the

[12] A description of this year's task can be found at: https://www2.statmt.org/wmt24/chat-task.html

limitations of current evaluation metrics in context-aware MT assessment. We advocate for the inclusion of our contextual annotation framework in QA processes to more accurately reflect the importance of context in MT workflows.

A full description of this work can be found in our paper (Menezes et al., 2023).

3.2.6 Dialogue Quality and Emotion Annotations for Customer Support Conversations

While NLP methods that are tailored to dialogue setups are becoming increasingly important within the NLP community, their advancement frequently hinders on the limited availability of annotated data. Specifically, the majority of data available are collected in controlled environments, thus not reflecting genuine conversations.

To address this limitation, the following paper describes an annotation proposal for emotion and conversational quality performed over the MAIA dataset (Farinha et al., 2022), a collection of genuine bilingual customer support conversations, whose interactions are aided by a bidirectional MT system. The work developed provides a unique and valuable resource for the development of text classification models. To this end, the authors present benchmarks for Emotion Recognition and Dialogue Quality Estimation.

The dataset was annotated along three dimensions:

1. Sentence level: corresponding to a single message;
2. Turn level: one or more sentences sent by one of the participants within a given time frame.
3. Dialogue level: a succession of turns between the customer and agent denoting the full conversation.

The metrics used for annotation are as follows:
For sentence Level Evaluation: • Correctness • Templated • Engagement
For turn level evaluation: • Understanding • Sensibleness • Politeness • Interaction Quality
For dialogue level evaluation: • Dropped Conversation • Task Success

In terms of emotions distribution along the MAIA dataset, the annotation outputs are shown in Figure 11.

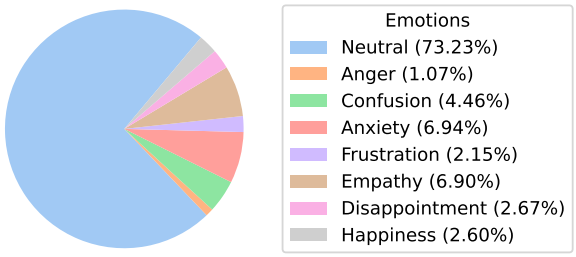

Figure 11: Emotion distribution of the MAIA dataset.

Regarding the development of the proposed benchmark, a fine-tuned XLM-RoBERTa was employed, following a train/dev/test splits at the dialogue level with a 70%/10%/20% distribution,

respectively. Performance was evaluated using Macro, Micro and individual emotion label F1 scores across all languages and the whole dataset.

We also fine-tuned a XLM-RoBERTa for Engagement using the ENDEX data (Xu et al., 2022); as well as Valid Sentence Prediction and Next Sentence Prediction using self-supervised data generated from DailyDialog (Li et al., 2017). Overall, we observed that depending on the language and target different models would demonstrate better performance.

Overall, this work presents a comprehensive emotion and dialogue quality annotation for the MAIA dataset, providing a novel opportunity to benchmark and explore applications of existing and future NLP models applied to dialogue. Results on the different benchmarks indicate there is still room for improving existing models.

A full description of this work can be found in our paper (Mendonça et al., 2023).

3.2.7 Findings of the WMT 2022 Shared Task on Chat Translation

In this the second edition of the Chat Translation Shared Task, we focus on translating bilingual customer support conversational text. Unlike the previous edition that used synthetic bilingual data, this year we used a segment of Unbabel's MAIA corpus, consisting of genuine bilingual conversations between agents and customers. We expanded the language pairs to include English-German (en-de), English-French (en-fr), and English-Brazilian Portuguese (en-pt-br), aiming to translate these conversations in a way that reflects their dynamic and informal nature, often rife with abbreviations, emoticons, and various errors.

The evaluation was carried out using both automatic metrics and human judgments through Multidimensional Quality Metrics (MQM), with the official ranking based on the overall MQM scores of participating systems in both directions. A total of 18 submissions from four different teams were received, with all teams participating in the English-German direction, and one team also taking part in the English-French and English-Brazilian Portuguese directions.

Our contributions can be summarised as follows:

- Demonstrating the feasibility and value of using genuine bilingual conversations for translation tasks, which can significantly impact the development of more nuanced and effective MT systems.
- Highlighting the importance of considering the unique challenges posed by informal, on-the-fly text production in customer support settings, such as the use of non-standard language, abbreviations, and emoticons.
- Providing a comprehensive evaluation framework that combines both automated metrics and detailed human evaluation, offering insights into the performance of MT systems in handling real-world conversational data.
- Releasing the MAIA Dataset to the research community, facilitating further exploration and innovation in conversational machine translation.

This work highlights the complexity of translating customer support chats and the need for MT systems to adapt to the informal and dynamic nature of such conversations. By focusing on genuine bilingual data and expanding language pairs, we contribute to a deeper understanding of the

challenges and opportunities in this area, paving the way for future advancements in machine translation technology. We aspire to continue for the third edition of this task within 2024, providing new, unseen data and updated annotations.

More information on the second edition of the task can be found in (Farinha et al., 2022).

Plans for future work

As LLMs are becoming ubiquitous in the field, we are planning to study further the use of context by these models and propose novel methods that better take context into account to improve performance in generation tasks. Furthermore, alongside the planned third edition of the chat translation shared task, we are aiming to further research methodologies that target translation in dialogue setups and can account for discourse and emotion-specific context.

4 Task T4.3: Simultaneous translation (UEDIN*, NAV)

Proposal

To be useful in dialogue, translation must be produced before the other party has finished their turn. This applies whether the modality is text or speech. However standard MT requires a full sentence before it can translate, often giving poor results when passed incomplete sentences. What we need is simultaneous translation, where we balance quality with latency (not making the user wait excessively) and possibly flicker (too much rewriting can give a bad user experience). Simultaneous translation can be implemented using retranslation or streaming. In the former approach, the system translates from the beginning of the sentence each time. In the latter approach, the system maintains a partial state and must decide between reading more source words and producing translation output. In the streaming approach, the system does not update already produced output, so it must be careful not to commit too early, whilst, in the retranslation approach, the system could create instability (flicker) if it constantly rewrites the translation. First approaches for end-to-end simultaneous speech translation were also very recently introduced as an IWSLT task. Among the possible improvements, using large LMs to predict what comes next and proposing decoding techniques that take into account this prediction uncertainty (as simultaneous interpreters would do), are interesting directions to improve simultaneous translation. As far as speech translation is concerned, translating a continuous and unsegmented input is another important challenge. Reinforcement learning for this task should be further investigated in order to find optimal read / write policies. Finally, using new decoder architectures where a dual decoder jointly transcribes and translates (the transcriber being ahead of the translator) is another interesting direction for simultaneous speech translation.

Proposal highlights

- Balance latency and flicker
- Simultaneous translation task IWSLT
- Speech translation and reinforcement learning

Completed Work

While this task was planned to start in the second half of the project, following advances in work for text and speech translation, we are already reporting some early work with findings on flicker reduction. Additionally, we describe the IWSLT shared task, for which UTTER partners were involved in the co-organisation, and which is highly relevant to this task, since it runs a collection of tasks related to speech translation, promoting further research and advancements in the field.

4.1 Self-training Reduces Flicker in Retranslation-based Simultaneous Translation

The aim of a Simultaneous speech translation (SimulST) system is to translate speech into text whilst obtaining the best possible tradeoff between quality and delay. Translation should normally be produced before the current sentence has been completed. SimulST is relevant for the translation of live speech (for example when subtitling a lecture) or translating a conversation. One approach to SimulST is retranslation, where the current sentence is retranslated from the beginning each time a new source word is produced. Although retranslation is a simple approach, requiring no modification of the underlying inference engine, it has the disadvantage that translation updates can cause potentially annoying flicker.

In this work, we propose a method for reducing flicker in retranslation-based systems. We apply self-training, in other words, we retrain the model on its own output, as a way of encouraging more monotonic output. Since the translation is more monotonic, translation updates are less likely to change the text that has already been output. This enables us to reduce flicker whilst maintaining similar translation quality to the original system. Our method can be combined with biased beam search (Arivazhagan et al., 2020) to further improve the latency-flicker tradeoff.

A full description of this work can be found in our paper (Sen et al., 2023).

4.2 IWSLT 2023 Shared Task on Simultaneous Speech Translation

Each year IWSLT runs a collection of tasks related to speech translation (ST). We were involved in the human evaluation of one of those tasks in 2023, the shared task on simultaneous speech translation. In this task, participants had to provide systems which operated with specified latency values on a development set and were judged on latency and translation quality on the test set. The evaluation protocol for this task involves replaying the source speech to annotators whilst the system output in the target language is displayed as subtitles. The annotators make regular judgements of translation quality using a 5-point scale. We can then average those judgements across all audio segments to provide an overall quality score for the system.

Full details of the evaluation are in the IWSLT findings paper (Agarwal et al., 2023).

Plans for future work

In the next steps, we aim to include speech generation so that the model is more adaptable to spoken language conversation. Some preliminary methods include Discrete Speech Units (Lam et al., 2024; Kim et al., 2024) and diffusion modelling (Zhu et al., 2023).

5 Conclusion

WP4 saw progress on all three tasks, with emphasis on Tasks 4.1 and 4.2 which started earlier in our timeline (following the plan and timeline of the UTTER proposal). Work concluded so far has led to contributions of datasets, methodology, software and empirical observations to advance adaptable and context-aware generation models, with emphasis on machine translation tasks. For the second half of the project, we aim to continue working across the three tasks, with anticipated steady progress and no foreseen risks for the continuation of the project.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.1. URL https://aclanthology.org/2023.iwslt-1.1.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and André FT Martins. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, 2023. URL https://aclanthology.org/2023.findings-emnlp.744/.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*, 2024. URL https://arxiv.org/abs/2402.17733.

Naveen Arivazhagan, Colin Cherry, Te I, Wolfgang Macherey, Pallavi Baljekar, and George Foster. Re-Translation Strategies For Long Form, Simultaneous, Spoken Language Translation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, December 2020. URL https://arxiv.org/abs/1912.03393.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*, 2015.

Ankur Bapna and Orhan Firat. Non-Parametric Adaptation for Neural Machine Translation. In *Proc. NAACL*, 2019.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. Building machine translation systems for the next thousand languages. 2022.

Christos Baziotis, Biao Zhang, Alexandra Birch, and Barry Haddow. When does monolingual data help multilingual translation: The role of domain and model scale, 2023. URL https://arxiv.org/abs/2305.14124.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. Findings of the WMT 2023 shared task on quality estimation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.52. URL https://aclanthology.org/2023.wmt-1.52.

Terra Blevins and Luke Zettlemoyer. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Nikolay Bogoychev and Pinzhen Chen. Terminology-aware translation with constrained decoding and large language model prompting. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.80. URL https://aclanthology.org/2023.wmt-1.80.

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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.298. URL https://aclanthology.org/2022.acl-long.298.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.522. URL https://doi.org/10.18653/v1/2021.emnlp-main.522.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/D07-1007.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H Wallach, H Larochelle, A Beygelzimer, F d\Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics, 2022a. doi: 10.18653/V1/2022.ACL-LONG.581. URL https://doi.org/10.18653/v1/2022.acl-long.581.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics, 2022b. doi: 10.18653/V1/2022.ACL-LONG.581. URL https://doi.org/10.18653/v1/2022.acl-long.581.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398.

Denis Emelin, Ivan Titov, and Rico Sennrich. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.616. URL https://aclanthology.org/2020.emnlp-main.616.

Ana C Farinha, M Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José GC De Souza, Helena Moniz, and André FT Martins. Findings of the wmt 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, 2022. URL https://aclanthology.org/2022.wmt-1.70/.

António Farinhas, José de Souza, and Andre Martins. An empirical study of translation hypothesis ensembling with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.733. URL https://aclanthology.org/2023.emnlp-main.733.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André FT Martins. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, 2021. URL https://aclanthology.org/2021.acl-long.505/.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.100. URL https://aclanthology.org/2022.naacl-main.100.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668, 2023a. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00626/118795.

Patrick Fernandes, Kayo Yin, Emmy Liu, André FT Martins, and Graham Neubig. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, 2023b. URL https://aclanthology.org/2023.acl-long.36/.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 2021.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022a. doi: 10.1162/tacl_a_00491. URL https://aclanthology.org/2022.tacl-1.47.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*,

pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.2.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.51. URL https://aclanthology.org/2023.wmt-1.51.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. Search engine guided neural machine translation. In *Proc. AAAI*, 2018.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 12 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00615. URL https://doi.org/10.1162/tacl_a_00615.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Sumire Honda, Patrick Fernandes, and Chrysoula Zerva. Context-aware neural machine translation for english-japanese business scene dialogues. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 272–285, 2023. URL https://arxiv.org/abs/2311.11976.

Vivek Iyer, Edoardo Barba, Alexandra Birch, Jeff Pan, and Roberto Navigli. Code-switching with word senses for pretraining in neural machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12889–12901, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.859. URL https://aclanthology.org/2023.findings-emnlp.859.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. Towards effective disambiguation for machine translation with large language models. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.44. URL https://aclanthology.org/2023.wmt-1.44.

Vivek Iyer, Arturo Oncevay, and Alexandra Birch. Exploring enhanced code-switched noising for pretraining in neural machine translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 954–968, 2023c. URL https://aclanthology.org/2023.findings-eacl.72.

Qingnan Jiang, Mingxuan Wang, Jun Cao, Shanbo Cheng, Shujian Huang, and Lei Li. Learning Kernel-Smoothed Machine Translation with Retrieved Examples. In *Proc. EMNLP*, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *Proc. ICLR*, 2021.

Minsu Kim, Jee-weon Jung, Hyeongseop Rha, Soumi Maiti, Siddhant Arora, Xuankai Chang, Shinji Watanabe, and Yong Man Ro. Tmt: Tri-modal translation between speech, image, and text by processing different modalities as different languages. *arXiv preprint arXiv:2402.16021*, 2024.

Mateusz Klimaszewski, Piotr Andruszkiewicz, and Alexandra Birch. Is Modularity Transferable? a Case Study through the Lens of Knowledge Distillation, 2024. Accepted to the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.

Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, 2017.

Tsz Kin Lam, Alexandra Birch, and Barry Haddow. Compact speech translation models via discrete speech units pretraining, 2024.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In Roger Levy and Lucia Specia, editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 333–342. Association for Computational Linguistics, 2017. doi: 10.18653/V1/K17-1034. URL https://doi.org/10.18653/v1/K17-1034.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

Frederick Liu, Han Lu, and Graham Neubig. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1121. URL https://aclanthology.org/N18-1121.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 11 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00343. URL https://doi.org/10.1162/tacl_a_00343.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, 2014.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. Efficient Machine Translation Domain Adaptation. In *Proc. ACL 2022 Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, 2022a.

Pedro Henrique Martins, Zita Marinho, and André FT Martins. Chunk-based Nearest Neighbor Machine Translation. In *Proc. EMNLP*, 2022b.

Pedro Henrique Martins, João Alves, Tânia Vaz, Madalena Gonçalves, Beatriz Silva, Marianna Buchicchio, José GC de Souza, and André FT Martins. Empirical assessment of knn-mt for real-world translation scenarios. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 115–124, 2023. URL https://aclanthology.org/2023.eamt-1.12.

John Mendonça, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C Farinha, Helena Moniz, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. Dialogue quality and emotion annotations for customer support conversations. *arXiv preprint arXiv:2311.13910*, 2023. URL https://arxiv.org/abs/2311.13910.

Miguel Menezes, Amin Farajian, Lisbon Unbabel, Portugal Helena Moniz, and Joao Graça. A context-aware annotation framework for customer support live chat machine translation. *MT Summit 2023*, page 286, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=MkbcAHIYgyS.

Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. Fast Nearest Neighbor Machine Translation. 2021.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR, 2022b. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Wafaa Mohammed and Vlad Niculae. On measuring context utilization in document-level MT systems. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.113.

NLLB team, Marta Costa-jussa, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janicec Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Gonzalez, Prangthip Hansanti, and Jeff Wang. No language left behind: Scaling human-centered machine translation. 07 2022. doi: 10.48550/arXiv.2207.04672.

Proyag Pal and Kenneth Heafield. Cheating to identify hard problems for neural machine translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1620–1631, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.120. URL https://aclanthology.org/2023.findings-eacl.120.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.21. URL https://aclanthology.org/2021.acl-long.21.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. ACL*, 2002.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Ponti. Modular deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=z9EkXfvxta. Survey Certification.

Matt Post. A Call for Clarity in Reporting BLEU Scores. In *Proc. Third Conference on Machine Translation*, 2018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In *Proc. EMNLP*, 2020.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, 2022a. URL https://aclanthology.org/2022.wmt-1.52/.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. Cometkiwi:

Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, 2022b. URL https://aclanthology.org/2022.wmt-1.60/.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73. URL https://aclanthology.org/2023.wmt-1.73.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. Findings of the WMT 2023 shared task on machine translation with terminologies. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.54. URL https://aclanthology.org/2023.wmt-1.54.

Sukanta Sen, Rico Sennrich, Biao Zhang, and Barry Haddow. Self-training Reduces Flicker in Retranslation-based Simultaneous Translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3734–3744, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.270. URL https://aclanthology.org/2023.eacl-main.270.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1009.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*, 2022.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*, 2022.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proc. NeurIPS*, 2014.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. Cross-lingual knowledge editing in large language models. *CoRR*, abs/2309.08952, 2023a. doi: 10.48550/ARXIV.2309.08952. URL https://doi.org/10.48550/arXiv.2309.08952.

Weixuan Wang, Barry Haddow, and Alexandra Birch. Retrieval-augmented multilingual knowledge editing, 2023b. URL https://arxiv.org/abs/2312.13040.

Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. Assessing the reliability of large language model knowledge, 2023c. URL https://arxiv.org/abs/2310.09820.

Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nischal Reddy Chandra, and Nanyun Peng. Endex: Evaluation of dialogue engagingness at scale. *arXiv preprint arXiv:2210.12362*, 2022.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. Findings of the WMT 2022 shared task on quality estimation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.3.

Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhang23m.html.

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. Guiding Neural Machine Translation with Retrieved Translation Pieces. In *Proc. NAACL*, 2018.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? *CoRR*, abs/2305.12740, 2023. doi: 10.48550/ARXIV.2305.12740. URL https://doi.org/10.48550/arXiv.2305.12740.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. Adaptive Nearest Neighbor Machine Translation. 2021.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024a.

Giulio Zhou, Tsz Kin Lam, Alexandra Birch, and Barry Haddow. Prosody in cascade and direct speech-to-text translation: a case study on korean wh-phrases. In *Findings of EACL*, 2024b. URL https://arxiv.org/abs/2402.00632.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. Modifying memories in transformer models. *CoRR*, abs/2012.00363, 2020. URL https://arxiv.org/abs/2012.00363.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*, 2024.

Yongxin Zhu, Zhujin Gao, Xinyuan Zhou, Ye Zhongyi, and Linli Xu. DiffS2UT: A semantic preserving diffusion model for textless direct speech-to-speech translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11573–11583, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.709. URL https://aclanthology.org/2023.emnlp-main.709.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D17 First report on adaptable and context