Graph Representations for Machine Translation in Dialogue Settings

Lea Krause[1]*, Selene Baez Santamaria[1], Jan-Christoph Kalo[2]

[1]Vrije Universiteit Amsterdam, [2]Universiteit van Amsterdam
{l.krause, s.baezsantamaria}@vu.nl, j.c.kalo@uva.nl

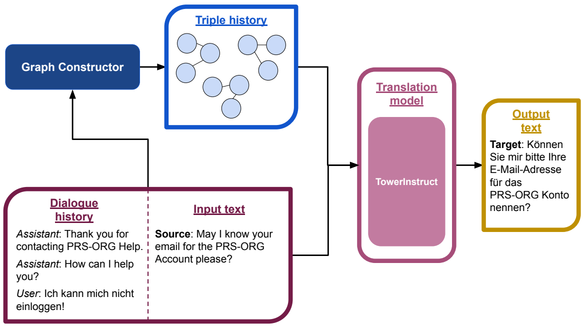
Abstract

In this paper, we present our approach to the WMT24 - Chat Task, addressing the challenge of translating chat conversations. Chat conversations are characterised by their informal, ungrammatical nature and strong reliance on context posing significant challenges for machine translation systems. To address these challenges, we augment large language models with explicit memory mechanisms designed to enhance coherence and consistency across dialogues. Specifically, we employ graph representations to capture and utilise dialogue context, leveraging concept connectivity as a compressed memory. Our approach ranked second place for Dutch and French, and third place for Portuguese and German, based on COMET-22 scores and human evaluation.

1 Introduction

Machine translation (MT) has been a prominent area of research, leading to the development of various approaches over the years (Maruf et al., 2021). While significant progress has been made, the majority of research has concentrated on refining methodologies rather than exploring the different types of text that require translation. A notable gap exists in the automatic translation of chat conversations—a gap that the WMT24 - Chat task specifically aims to address.

Chat conversations present unique challenges due to their informal, spontaneous nature, and frequent grammatical inconsistencies (Gonçalves et al., 2022). These characteristics starkly contrast with the more structured and formal text types, such as news articles, technical manuals, and political or medical documents, which have been the traditional focus of MT systems. In the context of chat translation, it is crucial to incorporate dialogue context effectively and to model the speakers and their language direction.



Assistant: Thank you for contacting PRS-ORG Help.

Assistant: How can I help you?

User: Ich kann mich nicht einloggen!

Source: May I know your email for the PRS-ORG Account please?

Target: Können Sie mir bitte Ihre E-Mail-Adresse für das PRS-ORG Konto nennen?

Figure 1: Approach 1: Triple-TowerInstruct

Recent advances in machine translation have increasingly leveraged large language models (LLMs). However, as noted by Maharana et al. (2024), LLMs often struggle with tasks requiring long-term memory, reasoning over historical context, and establishing long-range temporal or causal connections. These limitations are particularly problematic to dialogue tasks, where maintaining coherence and consistency across a conversation is vital.

To address these challenges, our system proposes enhancing LLMs with explicit memory mechanisms designed to support the generation of more consistent and coherent translations in dialogue settings[1]. We hypothesise that utilising graph representations will further improve the translation of chat conversations by capturing the connectivity between concepts, thus serving as a compressed memory of the dialogue context.

2 Related Work

In this section, we provide a brief overview of related work in the areas of conversational NLP, machine translation of conversational text, and text generation methods that incorporate knowledge graphs as an additional source of information.

* Corresponding author.

[1] All code and data related available at https://github.com/selBaez/chat-task-2024-data.

Proceedings of the Ninth Conference on Machine Translation, pages 1038–1046
November 15-16, 2024 ©2024 Association for Computational Linguistics

Conversational NLP Dialogue systems have a long-standing history in NLP. The advent of LLMs has led to significant improvements in the quality of these systems. However, a persistent challenge has been the limited context window of LMs, which restricts their ability to manage long chat histories effectively (Xu et al., 2021). To address this, retrieval-augmented models have been developed, which retrieve relevant passages from prior interactions to maintain coherence in dialogue over extended conversations (Xu et al., 2021). Recently, advancements in model architecture have resulted in substantially larger context windows, enabling state-of-the-art dialogue systems, such as ChatGPT, to operate effectively with this extensive LMs (Achiam et al., 2023).

Machine Translation Machine translation has seen remarkable advancements with the rise of large language models (Wang et al., 2023; Robinson et al., 2023). However, translating dialogues remains a particularly challenging task due to the informal and often context-dependent nature of conversational text (Gonçalves et al., 2022). The findings of recently shared tasks highlight ongoing difficulties and emerging solutions in this area (Farinha et al., 2022).

Our work is particularly related to the use of knowledge graphs in translation tasks (Moussallem et al., 2018; Zhao et al., 2021). In most existing approaches, multilingual knowledge graphs are leveraged to disambiguate and translate key entities within the text. This approach differs significantly from our method, as we employ a monolingual graph to store key information from the dialogue in a compressed format, facilitating more accurate and context-aware translations.

Graph-based Dialogue Systems Knowledge graphs have proven to be a valuable resource for grounding dialogue systems. The most common approach involves integrating large, external knowledge graphs to provide additional context and information that can enhance the dialogue's quality and relevance (Liu et al., 2019; Tuan et al., 2019; Zhang et al., 2020). While these approaches share a similar objective with our work, they fundamentally differ in that the knowledge graphs used are independent of the dialogue content itself.

In contrast, other approaches leverage graphs to represent the dialogue history, offering a structured way to maintain and utilise past interactions (Xu et al., 2020; Chen et al., 2023). This method enhances transparency, reduces the likelihood of hallucinations, and improves the system's ability to manage long-term conversations (Baez Santamaria et al., 2023). Our work aligns with this approach by utilising a graph to capture and organise key dialogue information, enabling more effective and contextually grounded dialogue systems.

3 Shared Task description

A dataset of original bilingual customer support conversations is provided. The language pairs available are English $\rightleftharpoons$ German (`en-de`), English $\rightleftharpoons$ Dutch (`en-nl`), English $\rightleftharpoons$ French (`en-fr`), English $\rightleftharpoons$ Brazilian Portuguese (`en-pt_br`), and English $\rightleftharpoons$ Korean (`en-ko`). Due to our team's language expertise, we decided to focus on the first four pairs.

4 System Overview

All our systems work with graphs extracted from dialogues. We employ a multi-step process to extract entities and relationships from the dialogue data and utilise these in various model settings. Our primary submission, **Triple-TowerInstruct**, integrates dialogue history into the translation process at inference, leveraging contextual cues to enhance performance across four language pairs. In addition to this, we explored an ablation study (TowerInstruct without dialogue history) and a novel model, **GraphFlanT5**, which combines graph and text embeddings within a unified framework.

4.1 Pre-processing

For generating the graphs, we perform entity and relation extraction by prompting GPT-4o. The prompt used for this process (see Prompt 1) is designed to extract relevant triples from the dialogue data, capturing the essence of interactions in a structured format. The system is instructed to analyse the dialogue and break it down into triples, each consisting of a subject, predicate, and object. These triples serve as the fundamental building blocks of the graph, representing the interactions between speakers.

In addition to extracting these triples, the prompt also instructs the system to annotate each triple with several attributes that provide deeper insights into the nature of the interactions. These annotations include:

- **Sentiment**: This attribute captures the emotional tone of the interaction, with values ranging from -1 for negative sentiment, 0 for neutral, and 1 for positive sentiment. This allows us to understand the emotional context in which the interaction takes place.

- **Polarity**: Polarity indicates whether the interaction involves a negation, affirmation, or is neutral or questioning. It is coded as -1 for negation, 0 for neutral or questioning, and 1 for affirmation. This helps in identifying the stance or intent behind the speaker's words and keeps the predicates uniform across negation, statements and questions (e.g. "don't travel" and "travel" receive the same predicate *travel* with different polarity scores)

- **Certainty**: This attribute is on a scale from 0 (uncertain) to 1 (certain), reflecting the speaker's confidence or the definitiveness of the statement. This helps in distinguishing between statements of fact and those that are speculative or uncertain and can subsequently be used by the model to communicate certainty about its knowledge more effectively.

- **Dialogue Act**: Dialogue acts categorise the type of speech act being performed, with predefined categories such as greeting, farewell, negative reaction, positive reaction, concern, query, and others.

4.2 Approach 1: Triple-TowerInstruct

In our first approach, we use the TowerInstruct-7B-v0.2[2] model, a variant of the Tower (Alves et al., 2024) family specifically designed for translation-related tasks.

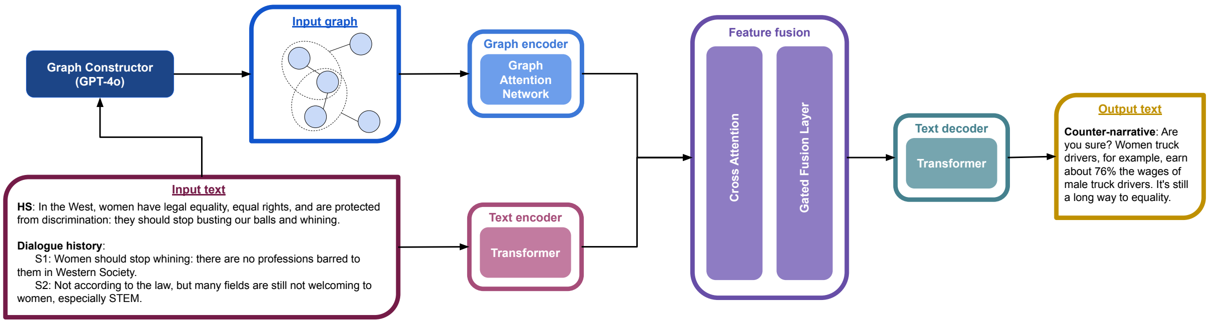
TowerInstruct-7B-v0.2 is based on the LLaMA-2 architecture, which has been extended through additional pretraining and fine-tuning to enhance its multilingual capabilities, outperforming other open models of similar scale. The model's foundation, TowerBase, was developed by continuing the pretraining on a diverse multilingual dataset across 10 languages (including Dutch, German, French, and Portuguese) incorporating both monolingual and parallel data to improve translation quality. Subsequently, TowerInstruct was fine-tuned using the TowerBlocks dataset, which includes a broad range of translation-related tasks and, relevant for the task of chat translation, multi-turn dialogue data from UltraChat (Ding et al., 2023). This fine-tuning process tailored the model specifically for translation workflows, making it adept at handling complex, multilingual interactions.

Prompt 1: Triple extraction with GPT-4o

```
system_prompt =
You will analyze a dialogue and break it down
into triples consisting of a subject, predicate,
and object. Each triple should capture the
essence of interactions between speakers.
Additionally, annotate each triple with:
- Sentiment (-1 for negative, 0 for neutral,
1 for positive)
- Polarity (-1 for negation, 0 for neutral/
questioning, 1 for affirmation)
- Certainty (a scale between 0 for uncertain
and 1 for certain)
- Dialogue act (
  0 : "greeting",
  1 : "farewell",
  2 : "negative_reaction",
  3 : "positive_reaction",
  4 : "concern",
  5 : "query",
  6 : "other")

Ensure that predicates are semantically
meaningful. Separate multi-word items with
an underscore.

Save it as a JSON with this format:
{
"Conversation ID": "60250de4b",
"dialogue": [
    {
      "sender": "customer",
      "text": "I can't find my order. It was
      supposed to arrive yesterday.",
      "triples": [
        {
          "subject": "I",
          "predicate": "cannot_find",
          "object": "my_order",
          "sentiment": -1,
          "polarity": -1,
          "certainty": 1,
          "dialogue_act": 4
        },
        {
          "subject": "It",
          "predicate": "was_supposed_to_arrive",
          "object": "yesterday",
          "sentiment": -1,
          "polarity": 1,
          "certainty": 0.7,
          "dialogue_act": 4
        }]},
    {
      "sender": "agent",
      "text": "I will help you with that.",
      "triples": [
        {
          "subject": "I",
          "predicate": "will_help",
          "object": "you_with_that",
          "sentiment": 1,
          "polarity": 1,
          "certainty": 1,
          "dialogue_act": 3
        }]}]}

user_prompt = f"Analyze the following con-
versation with ID {conversation_id}:
{conversation_text}"
```

Triple-TowerInstruct During inference, we merge the triple-based dialogue history, generated in the pre-processing stage (see Section 4.1), with

[2] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2.



HS: In the West, women have legal equality, equal rights, and are protected from discrimination: they should stop busting our balls and whining.

Dialogue history:
S1: Women should stop whining: there are no professions barred to them in Western Society.
S2: Not according to the law, but many fields are still not welcoming to women, especially STEM.

Counter-narrative: Are you sure? Women truck drivers, for example, earn about 76% the wages of male truck drivers. It's still a long way to equality.

Figure 2: Approach 2: GraphFlanT5

the current source sentence. This combined input, which encapsulates both the conversational context and the immediate translation task, is then fed into the model to ensure that the output accurately reflects the dialogue's ongoing flow and context.

As an ablation, we also evaluated the model's performance without providing dialogue history graphs.

4.3 Approach 2: GraphFlanT5

We have developed a unified framework named GraphFlanT5 (see Figure 2), which integrates both graph and text input into a single architecture, similar to (Yao et al., 2023). This model is designed to generate target sequences in text based on the dialogue history and the source sequence represented in text and graph forms.

As further preprocessing for this approach, we use spaCy's NeuralCoref[3] to resolve co-references, limiting the number of nodes to a maximum of 100. These are then represented as an adjacency matrix and fed into the main model.

To encode the graph, we employ a Graph Attention Network (GAT) (Veličković et al., 2018) with a single attention layer, followed by a dense layer and normalization. On the text side, we use a Transformer encoder for encoding. We specifically used FlanT5-base[4] for its multilingual capabilities. After obtaining the encoded features from both the graph and text, we apply cross-attention to align the text representation with the graph representation. A gated fusion mechanism (Wu et al., 2021) is then used to combine the outputs of the cross-attention. Finally, the fused features are passed into the Transformer decoder to generate the final textual answer.

We fine-tuned our model for 25 epochs with a learning rate of 5e-5 and a weight decay of 0.05. Training was conducted using mixed precision on two A10 GPUs.

5 Results & Discussion

5.1 Automated Metrics

Our primary submission, Triple-TowerInstruct, and its ablation variant without dialogue history graphs (NH) are compared against our second approach GraphFlanT5, the baseline (NLLB-200's (Team et al., 2022) 3.3B variant[5]), and the top-performing Unbabel system, using COMET-22 (Rei et al., 2022), Contextual-COMET-QE (Vernikos et al., 2022), BLEU (Papineni et al., 2002), and ChrF (Popović, 2015) scores [6].

Tables 1 and 2 show the results from our experiments across four language pairs: `en-de`, `en-nl`, `en-nl`, and `en-pt_br`. While we only submitted Approach 1 (Triple-TowerInstruct), we include the evaluation of the other approaches which were conducted after the shared task submission deadline. From the submitted approach, our team ranked second place for `en-nl` and `en-fr`, and third place for `en-pt_br` and `en-de` on the COMET-22 (Rei et al., 2020a) score.

Triple-TowerInstruct performed well across all language pairs, consistently outperforming the baseline based on COMET and in the majority of instances for the other metrics. For instance, in the `en-de` task, Triple-TowerInstruct achieved a COMET score of 91.3, outperforming the baseline's 89.8. The BLEU and ChrF scores further support this, with Triple-TowerInstruct scoring 53.0 in BLEU and 71.9 in ChrF for `en-de`, both above the baseline scores of 51.1 and 70.8, respectively. The

[3] https://github.com/huggingface/neuralcoref
[4] https://huggingface.co/google/flan-t5-base

[5] https://huggingface.co/facebook/nllb-200-3.3B
[6] Sacrebleu is used for the implementation of BLEU and ChrF (Post, 2018).

Model	en-de				en-nl			
	COMET	ChrF	BLEU	Context-COMET-QE	COMET	ChrF	BLEU	Context-COMET-QE
Triple-TowerInstruct	91.3	71.9	53.0	0.2039	90.9	70.6	48.0	0.0816
TowerInstruct NH	91.2	72.2	53.9	0.2128	91.3	66.2	44.7	0.1982
GraphFlanT5	85.3	65.1	44.5	0.0120	88.4	68.5	48.7	0.0697
Baseline	89.8	70.8	51.1	0.1730	88.1	62.6	38.7	0.0873
Unbabel+it	92.9	78.2	62.0	0.2526	93.6	79.8	63.9	0.1167

Table 1: Translation Results for German (en-de) and Dutch (en-nl). NH models refer to ablations without dialogue history. Results for the baseline and best performing system in the task (Unbabel+it) are included for comparison.

Model	en-fr				en-pt			
	COMET	ChrF	BLEU	Context-COMET-QE	COMET	ChrF	BLEU	Context-COMET-QE
Triple-TowerInstruct	91.6	75.7	58.8	0.0775	91.3	66.8	45.3	0.1909
TowerInstruct NH	91.7	75.2	57.9	0.0756	90.6	71.0	50.9	0.0686
GraphFlanT5	85.8	67.4	47.0	-0.1007	90.4	75.0	56.7	-0.0095
Baseline	90.1	76.2	58.7	0.0101	86.2	62.2	35.3	-0.0613
Unbabel+it	92.8	79.8	65.7	0.1034	93.9	79.7	65.0	0.2367

Table 2: Translation Results for French (en-fr) and Portuguese (en-pt). NH models refer to ablations without dialogue history. Results for the baseline and the best performing system in the task (Unbabel+it) are included for comparison.

NH variant, which omits dialogue history, saw a minor drop in performance for en-de and en-pt_br, with a drop in COMET score of 0.1 and 0.7 respectively, and slightly lower BLEU and ChrF scores. Interestingly, the opposite is true for the en-nl and en-nl language pairs. The Context-COMET-QE scores (Rei et al., 2020b), which are intended for reference-free machine translation evaluation and trained to reflect human judgements of the quality of translations, also demonstrated variability. For en-de, Triple-TowerInstruct scored 0.2039 in Context-COMET-QE (Rei et al., 2020b), while the NH variant scored 0.2128, showing a slight improvement when dialogue history was removed. While for en-pt_br including the history increased the score by 0.0383[7]. We also observed that COMET-based metrics and n-gram matching metrics (ChrF and BLEU) disagreed in ranking our TowerInstruct variants. When COMET favoured one variant, the n-gram metrics ranked it lower, and vice-versa. Underscoring the importance of using a combination of metrics, as relying on a single metric could give an incomplete picture of model performance.

GraphFlanT5 which integrates graph and text input within a unified framework, showed moderate results and did not outperform our TowerInstruct variants or the baseline in most cases. In the en-de task, GraphFlanT5 recorded a COMET score of 85.3, lower than both TowerInstruct and the baseline. Its BLEU and ChrF scores were also lower, at 44.5 and 65.1, respectively. However, in some tasks like en-nl, GraphFlanT5 performed competitively with a BLEU score of 48.7, suggesting that the integration of graph representations may offer benefits in certain contexts, but requires further optimisation to be competitive to more traditional approaches.

[7] See Kocmi et al. (2024) for an explanation of the different dynamic ranges of the mentioned metrics.

Model	en-de				en-nl			
	Formality	Lexical Cohesion	Pronouns	Verb Form	Formality	Lexical Cohesion	Pronouns	Verb Form
Triple-TowerInstruct	86.3	74.1	78.5	–	35.5	66.4	–	40.0
Baseline	79.4	76.0	79.1	–	53.0	57.4	–	35.7
Unbabel+it	88.6	82.9	70.5	–	93.9	87.7	–	54.5

Table 3: F1 Scores for German (en-de) and Dutch (en-nl) across different evaluation dimensions of MUDA. Where entries are left blank, the metric does not evaluate the language for that dimension.

Model	en-fr				en-pt			
	Formality	Lexical Cohesion	Pronouns	Verb Form	Formality	Lexical Cohesion	Pronouns	Verb Form
Triple-TowerInstruct	89.6	78.6	88.6	68.1	78.7	88.5	55.0	–
Baseline	86.9	82.1	82.0	70.2	45.7	81.0	55.8	–
Unbabel+it	91.3	90.2	92.9	74.2	88.0	95.5	74.4	–

Table 4: F1 Scores for French (en-fr) and Portuguese (en-pt) across different evaluation dimensions of MUDA.

5.1.1 MUDA

Tables 3 and 4 present the F1 scores for different evaluation dimensions—Formality, Lexical Cohesion, Pronouns, and Verb Form—of the Multilingual Discourse-Aware (MuDA) benchmark (Fernandes et al., 2023). We compared our primary model, Triple-TowerInstruct, against the baseline and the top-performing system, Unbabel+it. MuDA is designed to systematically evaluate machine translation models on their handling of discourse phenomena that require context. Unlike traditional metrics that focus broadly on translation accuracy, it specifically targets the model's ability to correctly translate discourse elements, such as pronouns and verb forms, that depend heavily on the surrounding context.

The performance of our model varied across different dimensions and language pairs, outperforming the baseline in 7 out of 13 cases. Overall, it demonstrated relatively strong performance on the **Formality** dimension, achieving competitive F1 scores in language pairs such as `en-de`, `en-nl`, and `en-pt_br`, with a notable increase of 33 points over the baseline for the latter. The exception was the `en-nl` pair, where the model's formality score was notably lower compared to both the baseline and top-performing systems, indicating a need for targeted improvements in handling formality specific to Dutch translations. However, performance on **Lexical Cohesion**, **Pronouns**, and **Verb Form** was less consistent across language pairs, with the model outperforming the baseline in only half of the cases.

5.2 Human Evaluation

Human evaluation confirms that our approach outperforms the baseline, and ranked second place for `en-nl` and `en-fr`, and third place for `en-pt_br` and `en-de` across all submitted approaches.

	en-de	en-nl	en-fr	en-pt
Triple-TowerInstruct	78.6	84.37	73.32	69.85
Baseline	74.5	53.07	67.81	56.37
Unbabel+it	84.22	92.22	79.62	78.0

Table 5: Human Evaluation Scores on document level for German (en-de), Dutch (en-nl), French (en-fr), and Portuguese (en-pt) across models.

The human evaluation was facilitated by the task organisers. It was conducted by professional linguists and translators using a combination of Direct Assessment and scalar quality metric (DA+SQM) implemented via the Appraise framework (Federmann, 2018).

6 Conclusion & Future Work

Our results underscore the importance of incorporating dialogue history in improving translation quality, highlighting its role in maintaining coherence and context throughout chat-based translations. The integration of graph-based representations also shows promise, particularly in capturing and leveraging the structural relationships within dialogue contexts. However, our findings indicate that further optimisation is required to fully realise the benefits of this approach, especially in terms of consistently outperforming more traditional text-based models.

In future work, one of our key objectives is to combine the strengths of TowerInstruct's translation capabilities with the advanced context modelling offered by our graph-based approach. By integrating these two methodologies, we aim to create a more robust system that can better handle the complexities of chat dialogue translation.

Furthermore, we plan to investigate the incorporation of additional contextual information, such as certainty or sentiment scores derived during preprocessing. These scores could potentially enhance the model's ability to weigh different parts of the dialogue based on their reliability and emotional tone, thereby improving overall translation accuracy. By factoring in sentiment, the model can better preserve the nuances of emotional expression within the conversation, leading to more contextually appropriate translations, which is particularly important in the task's customer service domain where frustration is common. By pursuing these directions, we aim to refine our models further, making them more adaptable and effective in real-world chat translation and dialogue tasks.

Acknowledgements

This research was funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) through the *Hybrid Intelligence Centre* via the Zwaartekracht grant (024.004.022), and the *Spinoza* grant (SPI 63-260) awarded to Piek Vossen.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. *Preprint*, arXiv:2402.17733.

Selene Baez Santamaria, Lea Krause, Lucia Donatelli, and Piek Vossen. 2023. The Role of Personal Perspectives in Open-Domain Dialogue: Towards Enhanced Data Modelling and Long-term Memory. *Proceedings of BNAIC/BeNeLearn the Joint International Scientific Conferences on AI and Machine Learning*, pages 1–19.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

Madalena Gonçalves, Marianna Buchicchio, Craig Stewart, Helena Moniz, and Alon Lavie. 2022. Agent and user-generated content and its impact on customer support MT. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 201–210, Ghent, Belgium. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792, Hong Kong, China. Association for Computational Linguistics.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *Preprint*, arXiv:2402.17753.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1–19.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *Preprint*, arXiv:1710.10903.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal 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 6153–6166, Online. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1835–1845.

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *Computing Research Repository*, arXiv:2305.16582.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2021. Knowledge graphs enhanced neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4039–4045.