Findings of the WMT 2022 Shared Task on Chat Translation

Ana C. Farinha[1*] **M. Amin Farajian**[1*]
Marianna Buchicchio[1] **Patrick Fernandes**[4,5,6] **José G. C. de Souza**[1]
Helena Moniz[1,2,3] **André F. T. Martins**[1,4,5]

[1]Unbabel, Lisbon, Portugal [2]INESC-ID, Lisbon, Portugal
[3]Faculdade de Letras, University of Lisbon, Portugal
[4]Instituto de Telecomunicações, Lisbon, Portugal
[5]Instituto Superior Técnico, University of Lisbon, Portugal
[6]Carnegie Mellon University, Pittsburgh, USA

Abstract

This paper reports the findings of the second edition of the Chat Translation Shared Task. Similarly to the previous WMT 2020 edition, the task consisted of translating bilingual customer support conversational text. However, unlike the previous edition, in which the bilingual data was created from a synthetic monolingual English corpus, this year we used a portion of the newly released Unbabel's MAIA corpus, which contains genuine bilingual conversations between agents and customers. We also expanded the language pairs to English↔German (en↔de), English↔French (en↔fr), and English↔Brazilian Portuguese (en↔pt-br).

Given that the main goal of the shared task is to translate bilingual conversations, participants were encouraged to train and test their models specifically for this environment. In total, we received 18 submissions from 4 different teams. All teams participated in both directions of en↔de. One of the teams also participated in en↔fr and en↔pt-br. We evaluated the submissions with automatic metrics as well as human judgments via Multidimensional Quality Metrics (MQM) on both directions. The official ranking of the systems is based on the overall MQM scores of the participating systems on both directions, i.e. *agent* and *customer*.

1 Introduction

With the significant translation quality improvements brought by newer machine translation (MT) approaches in the last years, we can start using MT to translate non-conventional content types such as bilingual and multilingual conversations. These new applications pose new challenges to MT systems and require new solutions to deal with them. In this shared task, we focus on the automatic translation of conversational text, in particular customer support chats, an important and challenging content due to their particular characteristics (Gonçalves et al., 2022; Wang et al., 2021; Farajian et al., 2020): In contrast to content types such as news articles and software manuals, among others, in which the text is carefully authored and well formatted, chat conversations are less planned, more informal, and often present ungrammatical linguistic structures. Furthermore, such conversations are usually on-the-fly production of text with very fuzzy frontiers with speech and mimicking speech production. Due to time requirements, since in dialogues turn-exchange need to be dynamic, the conversations may also have typos, abbreviations and ellipses. The conversations are also characterized by stressful moments, which in turn is represented by the capitalization of the entire word or turn, emoticons or emojis, and multiple punctuation marks.

Furthermore, Gonçalves et al. describe several factors that often lead to poor quality of the written text in this content type, resulting in lower quality of the MT outputs. They highlight the fact that the clients requiring customer support usually demonstrate high levels of impatience and frustration, resulting in typos, profanities, as well as variable capitalization and punctuation. They also mention that text is often times left unstructured, informal and agrammatical, factors that further increase the challenges of dealing with this particular content.

Given the limited number of parallel data for this domain, the main motivation for the Chat Translation Shared Task is to provide a common ground for evaluating and analyzing the challenges posed by conversational data as a content type, which has broad application in industry-level services. Following the success of the first edition of the Chat Translation Shared Task (Farajian et al., 2020), this

[*]These authors contributed equally.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 724 - 743
December 7-8, 2022 ©2022 Association for Computational Linguistics

customer	source_segment: Ola, tudo bem? target_segment: Hello! How are you?
customer	source_segment: Alguns meses atras, precisei restaurar o aplicativo da #PRS_ORG# para PC. target_segment: A few months ago, I needed to restore the #PRS_ORG# PC App.
customer	source_segment: Quando fiz isso, perdi todos os meus livros comprados. target_segment: When I did that, I lost all my purchased books.
customer	source_segment: Gostaria de saber como recupera-los. target_segment: I would like to know how to recover them.
customer	source_segment: Obrigada. target_segment: Thank you.
customer	source_segment: Celular para contato: #PHONENUMBER#. target_segment: Mobile for contact: #PHONENUMBER#
agent	source_segment: Thank you for the information. target_segment: Agradeço pela informação.
agent	source_segment: I will be more than happy to assist you. target_segment: Terei todo o prazer em ajudar você.
agent	source_segment: I see all your books are in the account linked to the #EMAIL# target_segment: Vejo que todos os seus livros estão na conta vinculada ao #EMAIL#

Table 1: Excerpt of a en↔pt-br conversation between a *customer* and an *agent*.

year we organized the second edition of the task with the following improvements:

- We released a genuine bilingual corpus, the Unbabel's MAIA Dataset. This consists of customer support dialogues in which the speakers (i.e. *customer* and *agent*) speak in their own language.

- We expanded the number of language pairs to three: English-German (en↔de), English-French (en↔fr), and English-Brazilian Portuguese (en↔pt-br).

- We performed manual evaluation on both directions of agent and customer, and we ranked the systems based on their overall performance in both directions, by using an adaptation of the multidimensional quality metrics (MQM) (Lommel et al., 2014) that is tailored to assess customer support translated content.

Similarly to the first edition of the task, we asked the participants to translate dialogues between two parties (i.e. *customer* and *agent*), where the *agent* writes in English and the *customer* writes in either German, French, or Brazilian Portuguese, depending on the language pair.

In order to evaluate the translation quality of the participating systems we used both automatic evaluation metrics and human judgement through MQM annotations. For the automatic evaluation metrics we used COMET (Rei et al., 2020), chrF (Popović, 2015), and SacreBLEU (Post, 2018), and for the human evaluation we used MQM (Lommel et al., 2014) performed with annotators specialized in explicit knowledge of translation errors and linguistics. Compared to the direct assessment evaluation (Graham et al., 2013, 2014, 2015) used in the previous edition, MQM annotations provide a more detailed analysis of the types and severities of the errors produced by the MT systems. MQM has also been shown to have a higher correlation with state-of-the-art metrics than direct assessments (Freitag et al., 2021).

This year, 18 submissions were received from 4 different teams, which have submitted outputs for both directions (i.e. *agent* and *customer*). Among these 4 teams, one team participated in all the three available language pairs, while the others focused only on en↔de. Details of their submissions and evaluation are described in §4 and §5.

2 The MAIA corpus

One of the biggest challenges of bilingual conversation translation, especially for Customer Support conversations, is the lack of appropriate publicly available datasets. To alleviate this issue, in the first edition of the Chat Translation Shared Task, Farajian et al. introduced the BConTrasT corpus that was based on a monolingual English corpus, Taskmaster-1 (Byrne et al., 2019). They translated the selected conversations into German mimicking a scenario in which an agent and a customer are communicating in their native languages. However, even this dataset was just an approximation of a real Customer Support conversation due to the fact that: 1) the original conversations in the Taskmaster-1 corpus were created by using crowdsourced workers interacting with each other to complete a specific task; and 2) the conversations were not truly genuine bilingual conversations since German segments were all just translations of the original English sentences.

This year, we made advancements by releasing the Unbabel's MAIA Dataset: a corpus that is truly composed of entire, genuine and original bilingual conversations from four different clients of the Unbabel database. The conversations are from clients that gave written consent on using this data for research purposes as long as in accordance with the General Data Protection Regulation (GDPR). In this corpus, the original segments of *customers* and *agents* are translated into their corresponding target languages via the MTPE process[1], done by the experienced translators of the Unbabel Community that have demonstrated consistently high quality within the respective language pair. MT segments were produced with a mixed of online MT services and internal ones. The corpus is released under the Creative Commons public license Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) and can be freely used for research purposes only. Please note that, as the license states, no commercial uses are permitted for this corpus. This data was collected within the MAIA Project (Martins et al., 2020).

The Unbabel's MAIA Dataset[2] contains more than 40k segments from more than 900 conversations in three language pairs (and a total of 6 language directions): English ↔ German (en↔de), English ↔ French (en↔fr) and English ↔ Portuguese (Brazil) (en↔pt-br). The breakdown of the corpus by language pair and direction is presented in Table 2. A sample conversation is presented in Table 1 and it shows that a conversation usually starts by the customer explaining the problem that led them to contact the *customer* support and is followed by the *agent* asking for more details in order to provide the necessary assistance.

Anonymization process. To make the conversations publicly available and in accordance with the General Data Protection Regulation (GDPR), we anonymized them first automatically by using the Unbabel proprietary anonymization tool and then by manually verifying the data. This resulted in 12 different anonymization categories, each presented by a specific token that are presented in Table 3. Importantly, Unbabel is also certified for ISO/IEC 27001:2013 Information Security Management Certification[3].

3 Task Description

Similarly to the first edition of the Chat Translation Shared Task, in this edition we focused on a critical challenge faced by international companies that are providing customer support in several different languages. One common approach to deal with this challenging requirement is centralizing the customer support with English speaking agents and having a translation layer in the middle to translate from the customer's language into the agent's language (e.g. English) and vice versa. The ideal solution for this environment needs to be able to properly handle all its aforementioned issues including the context-related challenges, the noisy inputs and multilingualism, among others.

In the second edition of the Chat Translation Shared Task we provide real genuine bilingual data for three different language pairs and encouraged the participants to make use of the bilingual context present in the conversations and to submit translations for both directions of the three language pairs. To emphasize on the importance of this aspect, we decided to rank the participating teams based on the overall quality of their primary submissions on both directions using a manual quality evaluation methodology through MQM annotations.

And, finally, we asked the participants to submit a maximum of three MT outputs per language pair direction, one primary and a maximum of two contrastive outputs. Due to time and budget constraints we performed the manual evaluation only for the primary submissions, while all the systems are evaluated using the automatic evaluation metrics (COMET, chrF, and SacreBLEU). For more details on the evaluation process please see §5.

3.1 Data

In the domain of customer support usually there is a very small amount of publicly available par-

[1] Machine Translation followed by a Post-Editing step.

[2] The full corpus can be downloaded from https://github.com/Unbabel/MAIA

[3] https://resources.unbabel.com/blog/unbabel-awarded-iso-iec-27001-2013-infor\mation-security-management-certification

MAIA corpus	en↔de	en↔fr	en↔pt-br
Number of conversations	496	264	164
Number of agent segments	8,509	9,911	4,741
Number of customer segments	9,468	5,115	3,674
Number of total (customer and agent) segments	17,977	15,026	8,415

Table 2: Number of conversations and segments in the MAIA corpus.

Token	Description
#NAME#	Person's names
#PRS_ORG#	Products, Services, and Organizations
#ADDRESS#	Address
#EMAIL#	E-mail address
#IP#	IP Address
#PASSWORD#	Password
#PHONENUMBER#	Phone number
#CREDITCARD#	Credit card number
#URL#	URL Address
#IBAN#	IBAN Address
#NUMBER#	Any number (all digits)
#ALPHANUMERIC_ID#	Any alphanumeric ID

Table 3: Anonymization tokens and their description.

allel data because of privacy and copyright issues that make releasing this kind of data difficult. In order to provide a more realistic setting, and due to the constraints outlined above, in this edition of the shared task, we provided participants with development and test sets only. The development sets can be divided into two types: SOURCE-ONLY, which that contains conversations without the human translations and PARALLEL, which that contains a smaller set of conversations with their corresponding human post-edited translations. The number of conversations and segments of the test and development sets per language pair and direction are presented in Table 4.

For training and validation purposes, participants were also allowed to use the training data of the general task (including the data of the previous editions), the data of the other tracks (eg. biomedical) and the other corpora (either parallel or monolingual) that are publicly available for research purposes including the data of the previous edition of the Chat Translation Task, BCONTRAST, as well as the corpora available on OPUS[4].

3.2 Baselines

In order to have a reasonable term of comparison for all the language direction, we used the large multilingual pre-trained M2M-100 model (Fan et al., 2021) with 418 million parameters. M2M is a multilingual MT model that supports all languages considered in this shared task. Moreover, since handling the context is one of the important challenges of the task we tried two baselines:

- A *sentence-level* baseline, where each utterance is passed separately to the model.
- A *context-level* baseline, where N consecutive utterances from the same conversation (and from the same direction) are passed and translated jointly by the model. In this paper we report the results of $N = 2$, that based on the automatic metrics performed the best on our validation sets.

While these models are not fine-tuned for *chat* conversation, they achieve good scores with automatic evaluation metrics and show the benefits of using context for this domain, even if they were not originally trained to use context.

We also report results of a larger version of the model (1B parameters) and different context sizes in Appendix A. In addition to these baselines, we also evaluated the results of four publicly available online MT systems on our test sets. In this paper we refer to them as Online-A, B, C, and D.

4 Participants

The participants were asked to submit at most three systems per language pair direction, one primary and a maximum of two contrastive ones. Moreover, the submitting team was required to explicitly indicate their primary and contrastive submissions. We received eighteen submissions from four different teams: BJTU-WeChat (two primaries and four contrastives), IITP-Flipkart (two primaries and four contrastives), HW-TSC (one primary and two contrastives), and Unbabel-IST (one primary and two contrastives). The first three teams participated only for en↔de, while the last team participated in all the language pairs and directions (i.e. en↔de, en↔fr, and en↔pt-br). Table 5 summarizes the participants and their affiliations.

[4] https://opus.nlpl.eu

	en↔de			en↔fr			en↔pt-br		
	source-only dev set	parallel dev set	parallel test set	source-only dev set	parallel dev set	parallel test set	source-only dev set	parallel dev set	parallel test set
Number of conversations	355	70	71	84	59	51	57	47	60
Number of total seg.	13,400	2,109	2,488	5,239	2,753	3,065	3,672	2,359	2,384
Number of agent seg.	6,389	1,006	1,113	3,305	1,750	1,937	2,007	1,353	1,381
Number of customer seg.	7,011	1,103	1,375	1,934	1,003	1,128	1,665	1,006	1,003

Table 4: Number of conversations and segments provided in the WMT 2022 Chat Translation Shared Task.

Team	Institution	Directions
BJTU-WeChat	Beijing Jiaotong University and WeChat	en↔de
HW-TSC	Huawei Translation Services Center	en↔de
IITP-Flipkart,	Indian Institute of Technology and Flipkart	en↔de
Unbabel-IST	Unbabel and Instituto Superior Técnico	en↔de, en↔fr, en↔pt-br
Online-A	A free publicly available online MT system	en↔de, en↔fr, en↔pt-br
Online-B	A free publicly available online MT system	en↔de, en↔fr, en↔pt-br
Online-C	A free publicly available online MT system	en↔de, en↔fr, en↔pt-br
Online-D	A free publicly available online MT system	en↔de, en↔fr, en↔pt-br

Table 5: The participating teams, their affiliations, and the directions that they participated.

All the participating systems follow a two step training in which a generic model is trained first on a large amount of publicly available data and then fine-tuned on the task data. The systems are different in the following aspects: i) the pre-training step, in which some use the publicly available models like mBART and Facebook-FAIR's WMT 2019, and the others train their own generic models, ii) the model architecture, in which some use deep encoder-decoder transformers, and others use multi-encoder transformers, iii) the fine-tuning stage and the data used in that step, and iv) the translation directions, in which some use bilingual models for each direction and some use a single multilingual model to cover all the language pairs and directions.

4.1 Systems

Here we briefly detail each participant's systems as described by the authors and refer the reader to the participant's submission for further details.

4.1.1 BJTU-WeChat

The joint submission of Beijing Jiaotong University and WeChat is an ensemble of deep Transformer models with 20 layers of encoder and 10 layers of decoder. Their models are firstly trained on the training corpora provided by the general track of WMT 2022. They are then fine-tuned on the training data of the chat translation track of WMT 2020 with several strategies to incorporate the potential context including the multi-encoder framework, speaker tag, and prompt-based fine-tuning.

Inspired by (Zhu et al., 2018) they proposed a Boosted Self-COMET-based Ensemble metric to evaluate the diversity of the generated hypotheses. As they report, it allows them to select some diverse, yet effective models from more than 100 models. Regarding the size of their models, the authors reports numbers that vary from 6.075 Billion to 6.881 Billion parameters.

4.1.2 IITP-Flipkart

IITP-Flipkart uses the Facebook-FAIR's WMT 2019 publicly available pre-trained models for en-de and de-en (Ng et al., 2019).[5] The models are based on the Transformer-big architecture (Vaswani et al., 2017). To fine-tune these models they follow a two-step procedure in which they first fine-tune the models on the training data of the Chat Translation track of WMT 2020 and then fine-tune the resulting models on the parallel validation set provided in the Chat Translation track of WMT 2022. To deal with the data scarcity issue of the task they combine the segments of agent and customer. To do so, for en-de, they use the agent subset of the above mentioned datasets as well as the customer segments by reversing their translation direction. The same applies to other direction.

[5] https://github.com/facebookresearch/fairseq/tree/main/examples/wmt19

For their primary submission they use a dual-encoder version of the WMT 2019 pre-trained FAIR model in which one encoder is used to encode the source context and the other one encodes the source segment. They use the weights of the encoder part of the pre-trained model to initialise the context-encoder weights. For the cross attention they use a weighted average of source-encoder and context-encoder attention. And for context they use the immediate previous source segment. Thus, the context can be either English or German, depending on the speaker of the previous utterance.

To analyze the impact of the context on the translation quality they experiment with a model that is trained with context and during inference it only uses the current sentence without any context. As they report, the results of this contrastive model confirm the observation of Li et al. that the improvement of the results are in some cases due to the fact that context acts as noise generator during training that makes the models more robust. And finally, their second contrastive model is a simple sentence level model that similarly as their primary model uses the Facebook-FAIR's WMT 2019 pre-trained model to initialise the weights. This model does not use any context.

As they reported, their primary submission is a model with 358 Million parameters. Their first contrastive model has the same number of parameters during training since it uses context, while during inference it uses only 312 million parameters since it does not use the context. This is the same number of parameters used by their second contrastive model that does not use any context at all.

4.1.3 HW-TSC

The Huawei Translation Services Center (HW-TSC)) team uses a deep transformer model with 25 layers of encoder and 6 layers of decoder. The model is pre-trained on the training data of the news track of WMT 2021. They used the bilingual validation set of the task to select in-domain data from the bilingual samples of the generic domain data. They reported the usage of self-training (i.e. forward translation), backward translation, model averaging, and context-aware translation.

4.1.4 Team Unbabel-IST

The joint submission of Unbabel and IST (Instituto Superior Técnico) uses the mBART50 model that has 12 layers of encoder and 12 layers of decoder. They fine-tuned the mBART50 model on a combination of the following two datasets: i) the in-domain parallel validation set, and ii) the samples similar to the validation set retrieved from the parallel generics corpus provided by the general track of WMT 2022. To find the similar samples they used LASER (Schwenk and Douze, 2017). At the inference time, their primary submission uses a retrieval-based approach in which for each segment of the test set the top-k nearest neighbors are retrieved from the following two data stores: i) the parallel in-domain validation set and ii) pool of the back-translated in-domain monolingual validation set of the task as well as the samples retrieved from the generic dataset that were used in the first stage of fine-tuning. Their first contrastive submission only uses the parallel validation set. Their second contrastive submission is the vanilla mBART50 model fine-tuned on the in-domain data, without the retrieval component.

Finally, concerning the model size, as they report it has the same number of parameters as mBART50, i.e. 761 million parameters.

5 Evaluation Procedures

Similarly to the previous edition, we evaluated the systems' performance both automatically and manually. This year we used COMET, chrF and SacreBLEU as the automatic metrics and MQM (Lommel et al., 2014) for the human evaluation. Due to time and budget constraints, the manual MQM evaluation was performed on the primary submissions only while all the submissions were evaluated using the automatic metrics. As mentioned earlier, the official rankings of the participating teams were based on the overall MQM score of their translations for the whole conversation, i.e. both *customer* and *agent* sides.

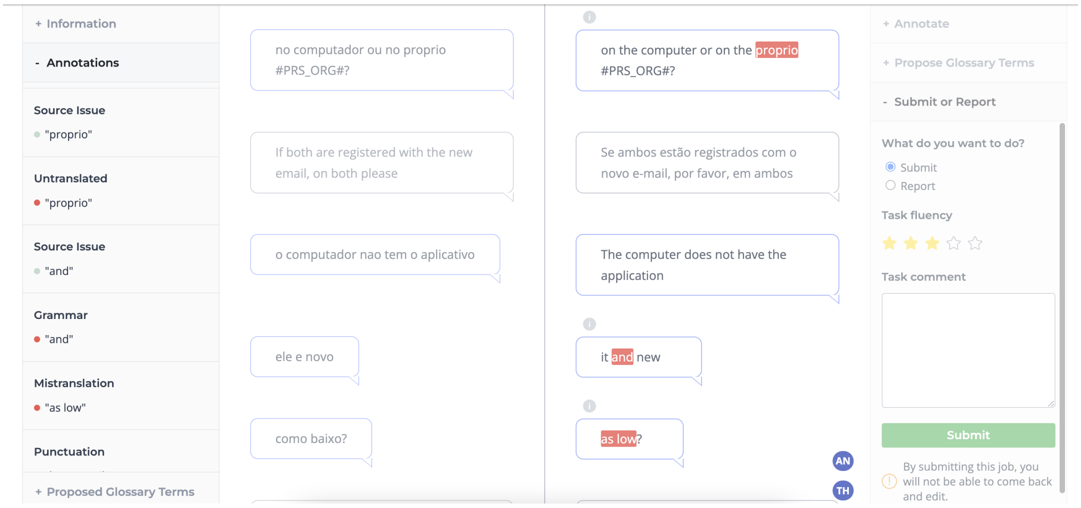
5.1 Automatic Evaluation

For the automatic evaluation of the systems' outputs we used COMET (`wmt20-comet-da`) (Rei et al., 2020), chrF (Popović, 2015), and SacreBLEU[6] (Post, 2018).

5.2 Human Evaluation

The human evaluation was performed by professional linguists and translators using an adaptation of the MQM framework (Lommel et al., 2014)

[6] We used version `2.1.0` with the signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0`



+ Information

- Annotations

Source Issue
"proprio"

Untranslated
● "proprio"

Source Issue
"and"

Grammar
● "and"

Mistranslation
● "as low"

Punctuation

+ Proposed Glossary Terms

no computador ou no proprio #PRS_ORG#?	on the computer or on the proprio #PRS_ORG#?
If both are registered with the new email, on both please	Se ambos estão registrados com o novo e-mail, por favor, em ambos
o computador nao tem o aplicativo	The computer does not have the application
ele e novo	it and new
como baixo?	as low?

+ Annotate

+ Propose Glossary Terms

- Submit or Report

What do you want to do?
● Submit
○ Report

Task fluency
★★★☆☆

Task comment

Submit

By submitting this job, you will not be able to come back and edit.

Figure 1: The human evaluation was performed on the Unbabel's proprietary Annotation Tool by showing the annotations the whole conversation. The figure refers to an excerpt of an en↔pt-br conversation annotation.

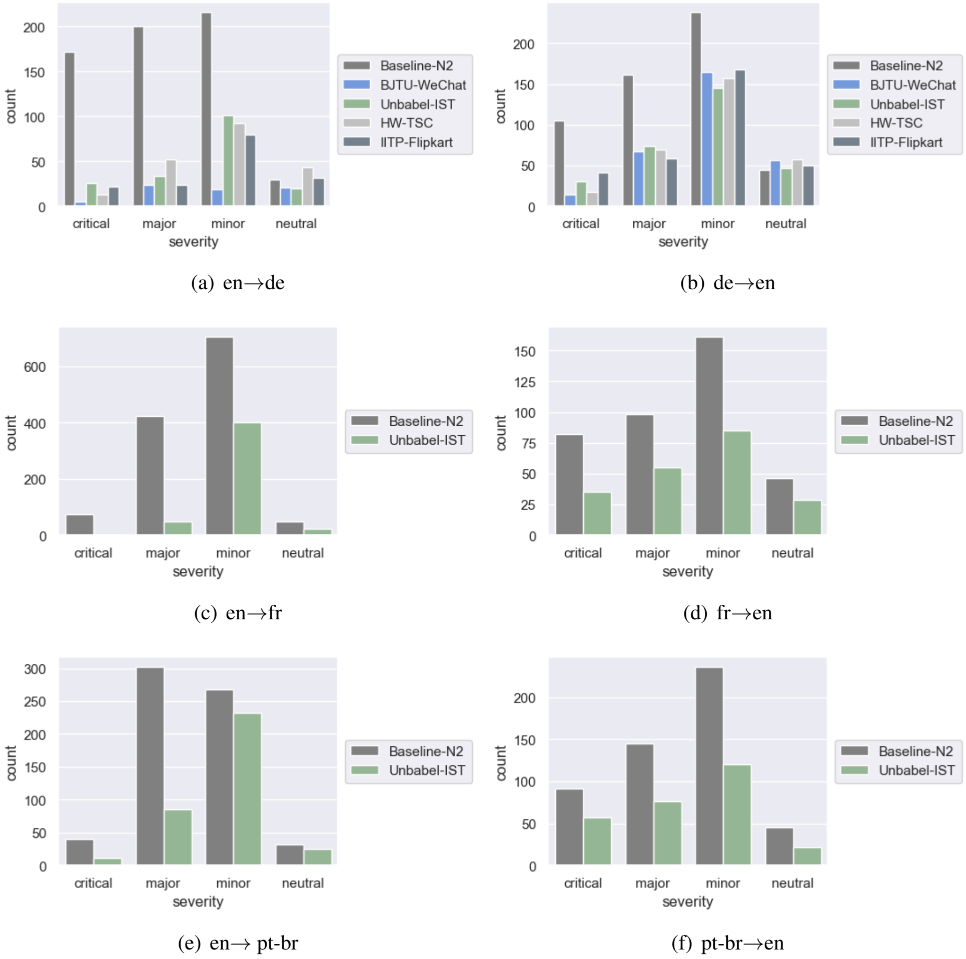
	en→de (agent)			de→en (customer)		
	COMET ↑	chrF ↑	BLEU ↑	COMET ↑	chrF ↑	BLEU ↑
Baselines						
Baseline without context	0.403	0.550	0.325	0.588	0.621	0.472
Baseline with context (N=2)	0.376	0.537	0.308	0.680	0.642	0.493
Primary submissions						
BJTU-WeChat	**0.810**	0.735	0.557	0.946	0.775	0.644
Unbabel-IST	0.774	0.733	0.557	0.915	0.737	0.612
IITP-Flipkart	0.768	0.730	0.549	0.907	0.729	0.582
HW-TSC	0.704	0.725	0.553	0.918	0.766	0.639
Contrastive submissions						
BJTU-WeChat, C1	0.804	0.731	0.551	0.948	0.780	0.646
BJTU-WeChat, C2	0.805	0.738	0.561	**0.951**	**0.778**	**0.648**
Unbabel-IST, C1	0.780	0.737	0.559	0.924	0.741	0.617
Unbabel-IST, C2	0.778	0.734	0.556	0.925	0.743	0.616
IITP-Flipkart, C1	0.769	0.730	0.550	0.905	0.729	0.582
IITP-Flipkart, C2	0.765	0.729	0.544	0.902	0.731	0.586
HW-TSC, C1	0.649	0.670	0.473	0.909	0.755	0.614
HW-TSC, C2	0.726	0.732	0.560	0.929	0.767	0.638
Online systems						
Online-A	0.758	**0.747**	**0.598**	0.903	0.733	0.579
Online-B	0.744	0.722	0.534	0.890	0.720	0.569
Online-C	0.717	0.707	0.515	0.877	0.730	0.575
Online-D	0.712	0.712	0.516	0.920	0.765	0.630

Table 6: Automatic metrics results for en↔de. The COMET scores are calculated with model `wmt20-comet-da`, and to calculate chrF and BLEU scores we used SacreBLEU.

that is tailored to assess Customer Support translated content (Gonçalves et al., 2022). The MQM-compliant typology used for this purpose is composed by:

- 8 parent categories, compliant with the newest version of the MQM framework[7]: *Accuracy*, *Linguistic Conventions*, *Terminology*, *Style*, *Locale Conventions*, *Audience Appropriateness*, *Design and Markup*, *Custom*;

[7] `https://themqm.info/typology/`



(a) en→de

(b) de→en

(c) en→fr

(d) fr→en

(e) en→ pt-br

(f) pt-br→en

Figure 2: Count of errors per severity for (a) en→de, (b) de→en, (c) en→fr, (d) fr→en, (e) en→pt-br, and (f) pt-br→en.

	en→fr (agent)			fr→en (customer)		
	COMET ↑	chrF ↑	BLEU ↑	COMET ↑	chrF ↑	BLEU ↑
Baselines						
Baseline without context	0.644	0.640	0.481	0.574	0.587	0.425
Baseline with context (N=2)	0.664	0.631	0.478	0.600	0.602	0.452
Primary						
Unbabel-IST	1.086	0.838	0.716	0.838	0.677	**0.544**
Contrastive						
Unbabel-IST, C1	1.082	0.836	0.712	0.840	0.676	0.542
Unbabel-IST, C2	**1.094**	**0.841**	0.718	**0.846**	0.675	0.542
Online systems						
Online-A	1.036	0.795	0.656	**0.846**	**0.678**	0.532
Online-B	1.085	0.838	**0.721**	0.827	0.669	0.517
Online-C	1.035	0.807	0.686	0.830	0.670	0.509
Online-D	1.044	0.788	0.618	0.819	0.673	0.521

Table 7: Automatic metrics results for en↔fr. The COMET scores are calculated with model `wmt20-comet-da`, and to calculate chrF and BLEU scores we used SacreBLEU.

	en→pt-br (agent)			pt-br→en (customer)		
	COMET ↑	chrF ↑	BLEU ↑	COMET ↑	chrF ↑	BLEU ↑
Baselines						
Baseline without context	0.824	0.681	0.495	0.610	0.631	0.471
Baseline with context (N=2)	0.863	0.675	0.493	0.675	0.653	0.496
Primary						
Unbabel-IST	1.077	0.771	0.621	0.849	0.689	0.547
Contrastive						
Unbabel-IST, C1	1.072	0.767	0.615	0.872	0.705	0.561
Unbabel-IST, C2	1.079	0.770	0.618	0.872	0.708	0.564
Online systems						
Online-A	0.965	0.725	0.551	**0.914**	**0.728**	**0.579**
Online-B	**1.084**	**0.791**	**0.647**	0.882	0.721	0.563
Online-C	1.069	**0.791**	0.643	0.887	0.726	0.559
Online-D	1.020	0.749	0.583	0.845	0.710	0.535

Table 8: Automatic metrics results for en↔pt-br. The COMET scores are calculated with model `wmt20-comet-da`, and to calculate chrF and BLEU scores we used SacreBLEU.

- 31 terminal nodes, including error types that are specific to MT, such as *MT Hallucination*[8] and customer support, such as *Source Issue*[9];
- 2 levels of granularity, composed by the 8 parent categories and the actual 31 terminal nodes (issue types) that annotators can use during the annotation process.

Regarding the severities attribution, we followed the same schema proposed in the MQM framework (Lommel et al., 2014), including a fourth severity, *Neutral*, to account for *Source Issue* errors. The definition of severities used in this evaluation are the following:

- *Neutral*: This severity degree is reserved only for the *Source Issue* category. This is used for linguistic issues that occur in the source text

[8] The *MT Hallucination* issue type is used when the MT generates a completely different translation that has no relation with the source text; the translation can still sound fluent and natural without reading the source, but the meaning is completely different. It is also used when the MT generates a chunk of repetitions in the target text or when the content is translated into gibberish: in other words, the machine generates an output made of non-words or repeated symbols.

[9] The *Source Issue* issue type needs to be used when there is an error in the target text and this is due to an issue in the source text. It can also be used when a part of the source text is written in the target language or in a different language, and the result is a mistranslation in the target.

	en↔de			en↔fr			en↔pt-br		
	agent	customer	overall	agent	customer	overall	agent	customer	overall
Baseline with context (N=2)	38.71	39.60	39.16	46.95	52.43	49.69	57.96	40.58	49.27
BJTU-WeChat	**96.44**	**80.09**	**88.27**	-	-		-	-	-
Huawei	88.33	79.02	83.68	-	-		-	-	-
Unbabel-IST	91.09	74.67	82.88	**90.08**	**77.21**	**83.65**	**84.16**	**69.01**	**76.59**
IITP-Flipkart	91.59	71.72	81.66	-	-		-	-	-

Table 9: MQM scores of the primary submissions of the participating teams, as well as the baseline MT systems.

that may have impact on the target translation and it is a signal of the overall quality of the source text to be translated;

- *Minor*: An error should be rated as minor if it does not lead to a loss of meaning and it does not confuse or mislead the user. It may, however, decrease the stylistic quality or fluency of the text, or make the content less appealing;
- *Major*: The usability or understandability of the content is impacted but it is still not unfit for purpose and the meaning of the content can be perceived as difficult to understand;
- *Critical*: The error severely changes the meaning of the original text. The reader cannot recover the actual meaning of the original text and the error carries health, safety, legal or financial implications to the end user/reader. In addition to this, a critical error also violates geopolitical usage guidelines, causes the application to crash or negatively modifies/misrepresents the functionality of the product or service. Finally, it can be offensive towards an individual or a group (a religion, race, gender, etc.).

To calculate the final MQM score per conversation the formula below is used:

$$\text{MQM} = 100 - \frac{I_{\text{Minor}} + 5 \times I_{\text{Major}} + 10 \times I_{\text{Crit.}}}{\text{Sentence Length} \times 100} \tag{1}$$

where I_{Minor} denotes the number of minor errors, I_{Major} the number of major errors and $I_{\text{Crit.}}$ the number of critical errors.

Figure 1 shows an excerpt of a pt-br customer conversation annotation, performed on a proprietary translation errors annotation tool from Unbabel. In this example, two *Source Issue* annotations are showcased, *proprio* and *and* that caused one Critical *Untranslated* error and one Critical *Grammar* error in the target text. In both cases, these examples outline some of the specificities of chat language and user-generated content, such as the lack of diacritics (Gonçalves et al., 2022) that can be observed in *proprio* and *e* on the source side (left pane) of the conversation. In the first case, *proprio*, a non-existent word in Portuguese, should have been written as *próprio*. As for the second case, *e* is a Portuguese conjunction that was translated literally into English, *and*, while the correct form should have been the verb *ser* (*to be* in English), conjugated in the 3rd person singular, *é*. The third error shown in Figure 1 shows yet another example of chat-specific language, such as the usage of a more idiomatic style (Gonçalves et al., 2022). The expression *como baixo?* refers to how *something can be downloaded from somewhere* and, besides its well-formedness in Portuguese, the style is idiomatic and conversational. The result is a *Mistranslation* error that refers to a literal translation into English. In this case, the meaning of the source text is completely lost and cannot be inferred by the reader.

Finally, as in the previous edition, we evaluated only a subsample of the full test set. For this, we randomly sampled conversations until the number of segments per direction was, at least, 500. We performed the annotations on both sides and calculated the overall conversation MQM score of each submission as the final score to use for the official ranking of the teams.

5.2.1 Customer Utility Analysis (CUA)

Besides reporting the overall MQM scores—the average MQM scores across conversations—, we decided to report, as a complement, a utility framework called Customer Utility Analysis (CUA) (Stewart et al., 2022). We decided to add this complementary analysis for two main reasons: 1) it gives us an idea of the quality across individual conversations; and 2) since MQM scores can be hardly understood without the context of a scale of reference or any direct connection the task-specific utility or value, CUA plots allow a better quality interpretation. This is possible because, as mentioned in §5.2, MQM is calculated by taking into account several factors, such as the total number of words of a conversation, the number of minor, major and critical errors and a severity multiplier. After the computation of the MQM scores at the conversation level, these are mapped into four different MQM buckets. In order to render this analysis more understandable from a visual point of view, we used a four color schema with the following MQM ranges:

- *Weak*: Dark Red (negative - 39 MQM)
- *Moderate*: Light Red (40 - 59 MQM)
- *Good*: Light Green (60 - 79 MQM)
- *Excellent*: Dark Green (80 - 100 MQM)

Ideally, we want the MT systems to have larger green and smaller red buckets, indicating less errors in the MT outputs and higher MQM scores.

6 Discussion

In this section we analyze the results of the automatic and human evaluation of the systems from different aspects.

6.1 Official ranking of the systems

The MQM scores of all the primary submissions as well the baselines (with context size 2) are presented in Table 9. As can be observed, in addition to the MQM score of each direction, the overall conversation-level MQM scores are also reported for each system.

Based on the overall MQM scores, the BJTU-WeChat team ranks first for the en↔de language pair, achieving higher MQM scores for both directions. This is consistent with the automatic scores of the systems reported in Table 6. BJTU-WeChat is followed by Huawei, Unbabel-IST, and IITP-Flipkart. As we can see in the distribution of the error severities in Figure 2, BJTU-WeChat produces significantly less critical and major errors in both directions. In the Neutral category we can see that all the systems perform almost the same, including the baselines. Based on this observation and the definition of this severity category (§5.2) we can infer that all the systems handle source-related issues more or less similarly. This calls for methods that are more reliable to source sentence issues, in particular for the *customer* side in which we have a significantly larger amount of issues when compared to the *agent* side.

By looking at the distribution of the error types presented in Table 15 we can see that "Mistranslation" is the most frequent error for all the systems. Given the definition of this error[10] and the fact that there was no in-domain training data for the given domain it was expected to see a large number of these errors in the outputs of all the MT systems.

For the en↔de language pair we received submissions from four teams, however, for en↔fr and en↔pt-br we received the outputs of one participating team only, making it more difficult to do an in-depth analysis on the results. The MQM scores of the baselines and the participating team are reported in Table 9, and their automatic scores can be found in Tables 7 and 8, respectively. As we can see, the Unbabel-IST systems outperform the baselines significantly both in terms of the manual MQM scores as well as the automatic metrics.

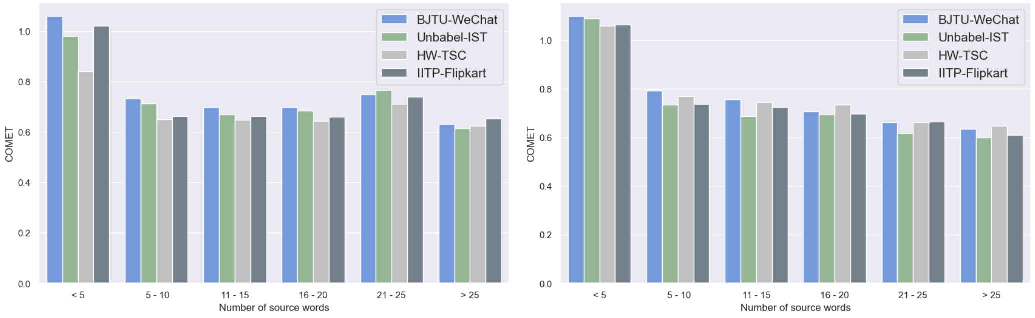
6.2 Computational efficiency of the approaches

The results of the primary submissions and the online systems (Table 6) shows that there is a big difference between the BJTU-WeChat submission and the other systems. As reported by the participants, this system is the only submission that uses an ensemble of a large number of models that makes it the least computationally efficient solution for the problem. The other submissions obtain results similar or better than the online systems and do not resort to model ensembling, making them more computationally efficient than the winning submission. The applicability and the computational efficiency of the models is one of the factors that we plan to pay more attention to in the future editions of the shared task.

6.3 Noisy source and its impact on MT quality

By comparing the MQM scores of the two directions (i.e. *agent* and *customer*) we can see that independently of the language pair, the scores of the *customer* side are significantly lower than the

[10] The word or phrase being translated wrongly according to the domain of interest.



(a) en→de

(b) de→en

Figure 3: COMET scores of the segments in different buckets based on the number of words in the source for (a) *agent* direction (en→de) and (b) *customer* direction (de→en).

scores of the *agent* side. This is in contrast to the previous observations that translating into English is usually easier than translating from English (Akhbardeh et al., 2021). We assume this is partially due to the different amounts of noise in the source segments of each direction and their impact on the final quality of the MT systems. To support our claim, we analyzed the source side of all the text conversations with a proprietary rule-based tool developed at Unbabel to detect spelling and grammatical errors, perform writing style checks (related to the formality of the text), among the detection of other types of issues that are specific to the content type of customer service conversations. As we can see in Table 10 the *customer* segments contain a larger degree of noise, up to 4 times, with respect to the *agent* side. We then proceeded by splitting the source segments of each direction into two sets of *noisy* and *non-noisy* categories and analyzed the quality of the models on each set separately. As we can see in Figure 5 the quality of the models on the noisy samples is significantly lower compared to the non-noisy samples. This is in line with the findings of Gonçalves et al., in which customers requiring customer support help usually exhibit high levels of impatience and frustration, that might be translated into agrammatical and unstructured text with lexical choices that often result in a degradation of the machine translation output.

	agent	customer
en↔de	55	105
en↔fr	40	116
en↔pt-br	24	95

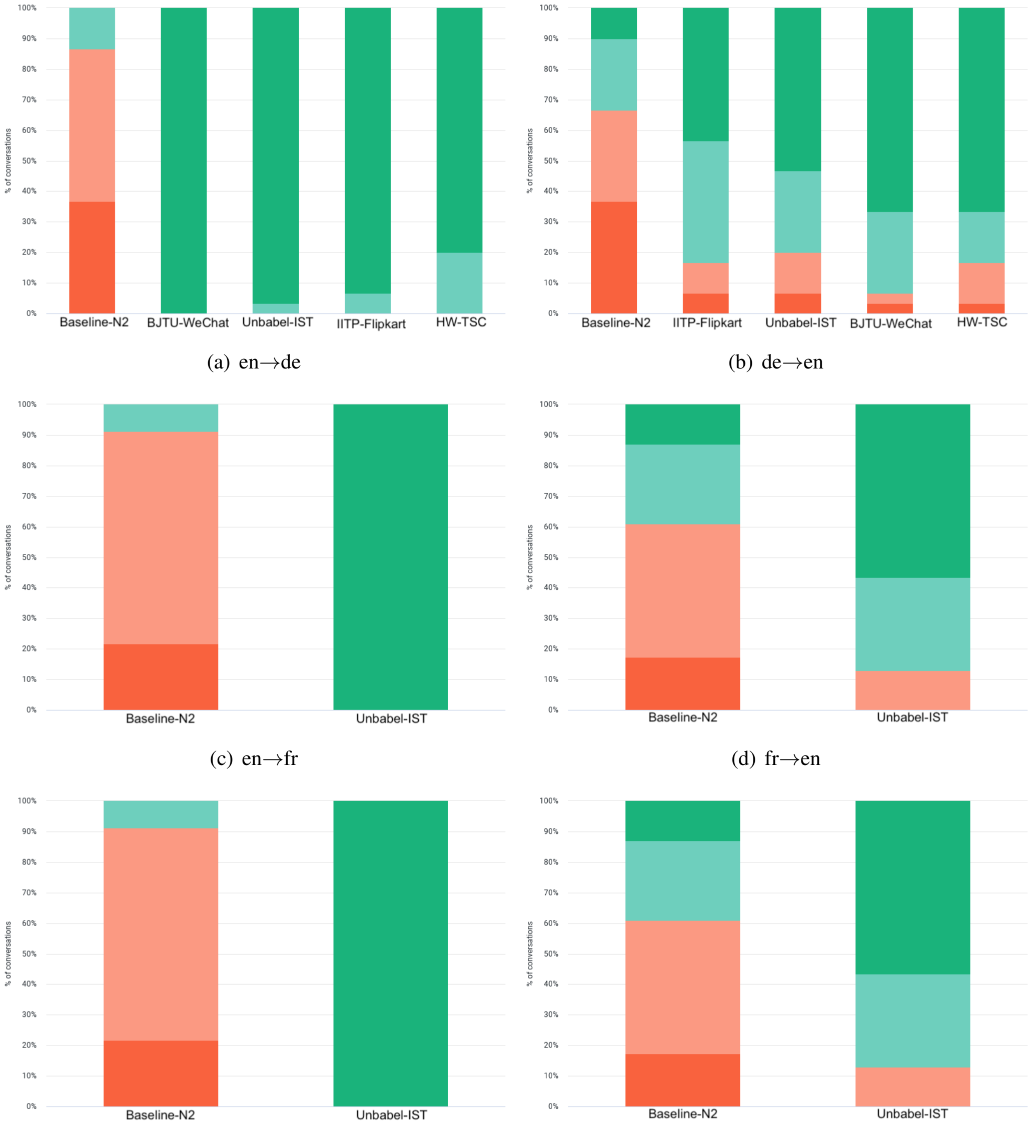
Table 10: The number of noisy source segments in each side of the test conversations.

6.4 Sentence length and MT quality

Looking at the test sets we can see varying lengths of source sentences, with the majority of them being very short segments (see Figure 6). To understand the impact of the sentence lengths on the final quality of the MT systems we grouped the input sentences into six buckets and measured the COMET score of each bucket (see Figure 3 for the COMET scores of the primary submissions of en→de and de→en). Independently of the direction, we can see that: 1) systems perform fairly similarly within each bucket; and 2) systems' performances tend to decrease as the number of source words increases. The pattern is very similar for the other language pairs and directions.

6.5 MT systems and CUA analysis

As mentioned in §5.2.1, CUA analysis provides complementary information to have a more clear understanding on the distribution of MQM scores. The bucketing approach used in CUA helps to easily interpret the quality of the MT systems. By looking into the de→en primary submissions, we can see that the BJTU-WeChat system not only outperforms the other systems significantly, but also produces the highest number of *excellent* translations. We can also see that, in general, the *agent* directions are easier to translate. In fact, no system produces any *Weak* or *Moderate* translations for this direction, while we can see a large number of *Weak* or *Moderate* ones in the outputs of all the systems for the *customer* direction.



(a) en→de

(b) de→en

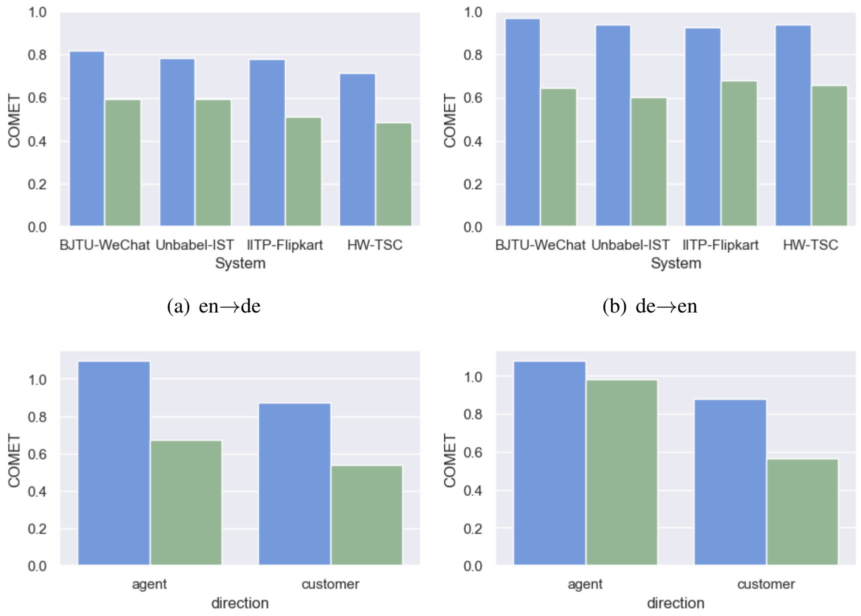
(c) en→fr

(d) fr→en

(e) en→pt-br

(f) ptbr→en

Figure 4: CUA plots for (a) en→de, (b) de→en, (c) en→fr, (d) fr→en, (e) en→ptbr, and (f) ptbr→en. Color schema: dark red (weak), light red (moderate), light green (good), and dark green (excellent).



(a) en→de

(b) de→en

(c) fr

(d) pt-br

Figure 5: COMET scores of the primary submissions on the *noisy* (in green) and *non-noisy* samples (in blue) of the test sets. The noises were detected by a proprietary tool developed at Unbabel. (a) shows the results on the agent direction (en→de), (b) shows the results on the customer direction (de→en), while (c) and (d) show the results of the only primary submissions (i.e. Unbabel-IST) for en↔fr and en↔pt-br, respectively.

6.6 Usage of context

All the four participating teams reported the incorporation of context in their experiments. But, depending on the approach, and the data they used as the context, they obtained different results and draw different conclusions. BJTU-WeChat used a simple prompt learning approach in which they add two preceding bilingual contexts at the tail of each utterance with a special token to indicate the boundary of the context. Their results show slight performance gains over the models that do not use the context. HW-TSC explores a similar approach but no promising results can be observed. This can be due to different factors like implementation details, the size and the combination of the data used as context, among other factors. For more details about the approaches and their difference please refer to their system description papers.

Differently than the BJTU-WeChat and the HW-TSC teams that use variations of the concatenation approach, IITP-Flipkart reports using an additional context encoder for incorporating context information. However, based on the test sets results we cannot observe any meaningful improvement over the system that does not incorporate the context, at least with automatic evaluation metrics.

Finally, Unbabel-IST report that in the few experiments they performed using context they did not observe any meaningful improvement on the performance of their models.

7 Conclusions

We presented the results of the WMT 2022 Chat Translation Shared Task. This year, we provided the participants with anonymized genuine bilingual Customer Support conversations for development and test sets. The conversations are part of the MAIA corpus, a corpus that we introduced here for the first time that aim to provide the best possible research ground for this very particular domain.

Four different teams participated in en↔de and one team participated also for en↔fr, and en↔pt-br. All participants covered both directions (i.e. *customer* and *agent*). We evaluated submissions with automatic metrics (i.e. COMET, chrF, and SacreBLEU) and primary submissions with MQM human evaluation. The MQM evaluations were conducted under an adaptation of the

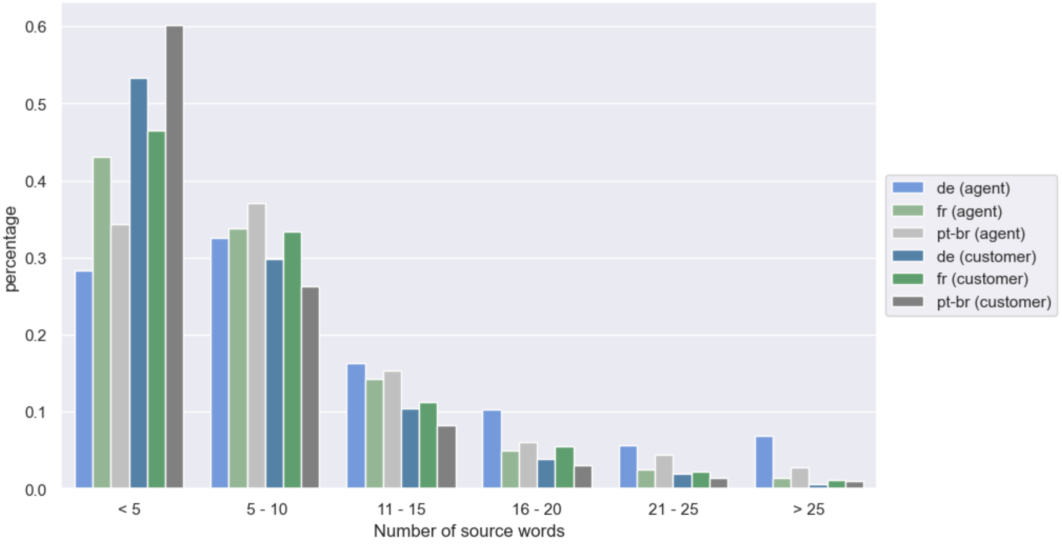


Figure 6: Percentage of segments when bucked according to the number of source words per lp and direction.

MQM framework (Lommel et al., 2014), that is tailored to assess Customer Support translated content (Gonçalves et al., 2022), providing a rich analysis of the type of errors that, we hope, will foster future MT research in this domain.

Acknowledgements

We would like to thank Vera Cabarrão and Miguel Menezes for their work in the anonymization of the data; Tânia Vaz, Madalena Gonçalves, and Maximilian Kohl for their help conducting the MQM annotations; Ricardo Rei and João Alves for useful discussions regarding the methodology used; and Pedro Gonçalves for the legal advise that made the release of the data possible. This work was supported by the P2020 programs MAIA (contract 045909), by the European Research Council (ERC StG DeepSPIN 758969), by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020 and by the European Union's Horizon Europe Research and Innovation Actions under Grant Agreement 101070631 (UTTER).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4506–4517.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021.

Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23:3 – 30.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

André F. T. Martins, Joao Graca, Paulo Dimas, Helena Moniz, and Graham Neubig. 2020. Project MAIA: Multilingual AI agent assistant. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 495–496, Lisboa, Portugal. European Association for Machine Translation.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. 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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. 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.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 157–167. Association for Computational Linguistics.

Craig A Stewart, Madalena Gonçalves, Marianna Buchicchio, and Alon Lavie. 2022. Business critical errors: A framework for adaptive quality feedback. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 231–256, Orlando, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30.

Tao Wang, Chengqi Zhao, Mingxuan Wang, Lei Li, and Deyi Xiong. 2021. Autocorrect in the process of translation — multi-task learning improves dialogue machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 105–112, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

A Baseline Results with Other Context Sizes

Table 11: M2M-418M results on the *development* set with various context sizes

Lang	Direction	Context Size	BLEU	ChrF	COMET
de	agent	all	31.94	53.25	0.2504
		0	**35.24**	**57.17**	**0.4168**
		1	33.80	56.05	0.3910
		2	33.89	55.96	0.3811
		3	33.40	55.76	0.3648
		5	33.07	55.23	0.3493
	customer	all	47.14	62.05	0.6114
		0	45.98	60.81	0.5426
		1	**48.28**	**62.80**	**0.6326**
		2	47.11	62.06	0.6163
		3	47.23	62.08	0.6073
		5	47.52	62.41	0.6225
fr	agent	all	45.67	61.02	0.5105
		0	54.14	69.47	0.7984
		1	**54.72**	69.83	**0.8173**
		2	53.58	68.81	0.7978
		3	53.68	69.00	0.7973
		5	52.69	68.00	0.7750
	customer	all	48.14	63.77	0.6784
		0	46.51	62.29	0.6382
		1	**48.35**	63.53	0.6526
		2	48.05	**63.61**	**0.6834**
		3	48.52	64.06	0.6786
		5	48.17	63.74	0.6753
pt-br	agent	all	45.60	63.25	0.7801
		0	**50.38**	**68.84**	0.8645
		1	49.67	67.95	**0.9129**
		2	49.94	67.95	0.9029
		3	49.11	67.41	0.9116
		5	48.67	66.95	0.8935
	customer	all	47.10	62.29	0.6449
		0	44.71	59.95	0.5851
		1	46.88	62.06	0.6332
		2	**47.24**	**62.31**	0.6437
		3	46.96	**62.31**	**0.6491**
		5	47.30	62.53	0.6514

Table 12: M2M-1B results on the *development* set with various context sizes

Lang	Direction	Context Size	BLEU	ChrF	COMET
de	agent	all	33.64	50.81	0.0070
		0	**43.36**	63.90	**0.4696**
		1	39.86	59.56	0.2938
		2	36.96	54.70	0.1669
		3	35.04	52.66	0.0926
		5	34.52	51.63	0.0505
	customer	all	49.84	64.41	0.5192
		0	**60.20**	74.03	**0.8307**
		1	59.44	72.44	0.7976
		2	57.08	70.71	0.7620
		3	57.87	71.52	0.7889
		5	57.18	70.73	0.7554
fr	agent	all	48.67	65.78	0.7100
		0	**55.16**	72.33	0.8718
		1	52.67	70.21	0.8857
		2	51.75	69.47	0.8873
		3	52.58	70.45	**0.8988**
		5	49.89	68.94	0.8122
	customer	all	50.02	64.56	0.6510
		0	50.33	64.73	0.6434
		1	50.18	64.79	**0.6626**
		2	**50.57**	65.21	0.6550
		3	50.24	64.86	0.6546
		5	50.31	64.86	0.6551
pt-br	agent	all	48.63	64.10	0.6409
		0	49.26	64.31	**0.6884**
		1	49.51	64.46	0.6362
		2	49.76	64.92	0.6449
		3	**49.79**	65.21	0.6597
		5	48.56	64.03	0.6422
	customer	all	48.48	63.51	0.6401
		0	45.99	61.18	**0.6427**
		1	**48.98**	63.89	0.6424
		2	48.22	62.57	0.6167
		3	48.30	63.17	0.6415
		5	48.25	63.13	0.6241

	en→fr (agent)		fr→en (customer)	
	Baseline-N2	Unbabel-IST	Baseline-N2	Unbabel-IST
Addition	39	17	37	12
Agreement	6	6	2	3
Capitalization	55	32	12	0
Currency Format	0	2	0	0
Date/Time Format	2	2	4	2
Grammar	90	24	36	22
MT Halucination	8	0	19	0
Mistranslation	469	60	113	83
Omission	21	11	42	28
Punctuation	170	98	18	4
Register	2	0	0	0
Source Issue	50	22	46	29
Spelling	19	20	2	0
Unnatural Flow	3	1	0	0
Untranslated	207	2	25	7
Whitespace	86	161	2	0
Word Order	26	18	16	4
Wrong Named Entity	0	0	13	10

Table 13: Counts per error type for fr .

	en→pt-br (agent)		pt-br→en (customer)	
	Baseline-N2	Unbabel-IST	Baseline-N2	Unbabel-IST
Addition	14	8	47	13
Agreement	9	1	0	1
Capitalization	24	18	27	19
Currency Format	3	2	0	0
Grammar	53	31	54	28
MT Halucination	1	0	28	6
Mistranslation	224	99	145	100
Omission	25	30	84	42
Punctuation	131	109	19	15
Register	2	0	0	0
Source Issue	32	25	46	23
Spelling	2	2	0	0
Unnatural Flow	8	2	0	0
Untranslated	97	5	22	4
Whitespace	4	5	8	4
Word Order	10	4	25	15
Wrong Language Variety	1	6	0	0
Wrong Named Entity	2	6	12	5

Table 14: Counts per error type for pt-br.

	en→de (agent)					de→en (customer)				
	Baseline-N2	BJTU-WeChat	Unbabel-IST	IITP-Flipkart	HW-TSC	Baseline-N2	BJTU-WeChat	Unbabel-IST	IITP-Flipkart	HW-TSC
Addition	19	1	7	4	7	38	10	23	14	19
Agreement	1	3	0	1	0	0	0	1	0	0
Capitalization	2	1	2	1	1	31	7	4	6	5
Currency Format	0	0	0	0	0	0	1	0	0	0
Date/Time Format	0	0	0	0	0	2	1	2	2	4
Grammar	82	1	35	13	22	46	25	32	30	23
Inconsistency	0	0	1	0	0	0	2	0	2	1
MT Halucination	3	0	0	0	1	18	0	4	0	0
Measurement Format	0	0	0	0	0	1	1	1	1	1
Mistranslation	258	34	58	51	81	179	66	84	56	72
Omission	81	2	11	11	10	87	77	60	85	78
Punctuation	6	0	4	10	10	28	19	9	17	15
Register	18	0	0	0	3	0	0	0	0	0
Source Issue	31	22	20	32	43	47	57	47	50	58
Spelling	0	3	0	1	2	1	0	0	1	0
Untranslated	81	0	19	15	6	15	9	3	5	4
Whitespace	7	0	14	14	4	12	7	5	30	3
Word Order	29	0	8	5	10	42	17	18	13	17
Wrong Named Entity	0	2	2	0	0	2	3	3	2	2

Table 15: Counts per error type for de.