Findings of the WMT 2020 Shared Task on Chat Translation

M. Amin Farajian[1*] **António V. Lopes**[1*] **André F. T. Martins**[1,3]
Sameen Maruf[2] **Gholamreza Haffari**[2]
[1]Unbabel, Rua Castilho 52, 1250-069, Lisbon, Portugal
[2]Monash University, VIC, Australia
[3]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal
{amin, antonio.lopes, andre.martins}@unbabel.com
{sameen.maruf, gholamreza.haffari}@monash.edu

Abstract

We report the results of the first edition of the WMT shared task on Chat Translation. The task consisted of translating bilingual conversational text, in particular customer support chats for the English-German language pair (English agent, German customer). This task varies from the other translation shared tasks, i.e. news and biomedical, mainly due to the fact that the conversations are bilingual, less planned, more informal, and often ungrammatical. Furthermore, such conversations are usually characterized by shorter and simpler sentences and contain more pronouns.

We received 14 submissions from 6 participating teams, all of them covering both directions, i.e. En→De for *agent* utterances and De→En for *customer* messages. We used automatic metrics (BLEU and TER) for evaluating the translations of both agent and customer messages and human document-level direct assessments to evaluate the agent translations.

1 Introduction

Despite the significant progress in Neural Machine Translation (NMT) in the last years (Vaswani et al., 2017; Hassan et al., 2018), most systems still operate at sentence-level, disregarding the context of previous sentences. It has been pointed out that ignoring the context may degrade the quality of translations, leading to incorrect choice of pronouns, lexical inconsistency, and incoherence (Läubli et al., 2018; Toral et al., 2018). This is particularly relevant in the context of bilingual chat translation, which normally consists of short messages, referencing each other, and where the correct lexical choice to translate a speaker might have been uttered in a previous turn by the other speaker.

Numerous systems have been proposed recently to address document-level translation (Tiedemann and Scherrer, 2017; Zhang et al., 2018; Maruf et al., 2019; Miculicich et al., 2018; Voita et al., 2019b; Tu et al., 2018; Maruf et al., 2018; Jean et al., 2017; Voita et al., 2018, 2019a; Junczys-Dowmunt, 2019; Lopes et al., 2020), focusing on extending both Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) with additional encoders or decoders to incorporate previous sentences context. However, often, the approaches are developed for single speaker and document-like tasks. By contrast, in this shared task, we focus on the online multispeaker and multi-lingual setting, where each participant in the conversation speaks in their native language. This task has been first considered by Maruf et al. (2018).

In the first round of the Chat Translation shared task, we propose translating dialogues with two speakers, where the first speaker is speaking in the German→English direction and the second is speaking in the English→German. Moreover, we tailor this task for a specific use case: translating conversational text of the customer support chats. In this setting the utterances of the German speaking customer are translated using a machine translation system into English. Then, the replies of the English speaking agent are translated into German and sent to the customer.

Translating conversational text, in particular customer support chats, is an important and challenging application task for machine translation technology. This type of content has so far not been extensively explored in prior MT research, largely due to the lack of publicly available data sets. Prior related work has mostly focused on movie subtitles and European Parliament speeches. To alleviate this problem, we created a corpus for this shared task, ***BConTrasT***(§2), which is translated from English into German and is based on the monolingual Taskmaster-1 corpus (Byrne et al., 2019).

*These authors contributed equally.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 65–75
Online, November 19–20, 2020.

The main motivation of this shared task is to analyze the challenges posed by conversational data as a content type, which has a broad application in industry-level services. In this content type, the text is usually not carefully well formatted, frequently contains typos, abbreviations, and inconsistent casing, usually with shorter sentences, often informal and ungrammatical. Since chat sessions are interactive, the task of translating conversations can be seen as a two-in-one task, modelling both dialogue and document-level translation at the same time.

In order to evaluate the translation quality of the participating systems we use both automatic metrics (BLEU (Papineni et al., 2002) and TER (Snover et al., 2006)), and human evaluation, consisting of Direct Assessment (DA). For DA, we define the evaluation process similarly to last year's WMT News Translation task (Barrault et al., 2019) with document-level context and following the set of recommendations of Läubli et al. (2020). However, differently than the News task, here we rely on professional translators instead of a crowd. This is mainly based on the observations of Läubli et al. (2020), which provides evidence of the professional translators having better judgment and ability to detect fine-grained phenomena.

Six teams participated in this first campaign of the Chat Translation shared task, with 14 runs in total. All teams submitted both English→German and German→English directions. In §4, we describe each system in more details.

2 Bilingual Conversational Data

One of the main challenges of bilingual conversation translation is the lack of publicly available data sets targeted for the task. The most commonly used datasets are movie subtitles (Lison and Tiedemann, 2016), European Parliament speeches (Koehn, 2005), and conversations extracted from the public forums such as Ubuntu Dialogue corpus (Lowe et al., 2015). These corpora, however, usually involve more than two speakers, contain a significant amount of noise (e.g. speakers information missing in the case of movie subtitles), and usually cover very broad domains.

For the Chat Translation task, we aim to develop a common ground for MT researchers to train and test their solutions by providing common training, validation, and test sets, as well as a common shared task definition. Unfortunately, due to the General Data Protection Regulation (GDPR), most commercial enterprises cannot distribute publicly their proprietary data. Therefore, we opted for using the Taskmaster-1 corpus (Byrne et al., 2019), which includes monolingual (English) task-based dialogues in six domains: (i) ordering pizza, (ii) creating auto repair appointments, (iii) setting up ride service, (iv) ordering movie tickets, (v) ordering coffee drinks, and (vi) making restaurant reservations. We used this corpus for creating the data of our shared task.

Since the main goal of this task is to enable multilingual speakers communicate with each other in their native language, we used the Unbabel translation service [1] to translate the utterances of both speakers into the target language (German). In this process, the conversations (originally in English) were first automatically translated into German and then manually post-edited by Unbabel editors, who are native German speakers. Having the conversations in both languages allows us to simulate bilingual conversations in which one speaker, the *customer*, speaks in German and the other speaker, the *agent*, answers in English. Table 1 shows the first few sentences of a bilingual conversation, along with their corresponding translations. In order to provide a realistic environment in which the amount of in-domain parallel data is scarce, we translated only a small set of the Taskmaster-1 corpus. Since pronouns are one of the main challenges in translating conversational data, we selected the conversations that contain at least one English anaphoric pronoun *it*. For this we used NEURALCOREF [2] and selected around 18k sentence pairs and then divided them into train, development, and test sets (see Table 2).

3 Task Description

A critical challenge faced by international companies today is delivering customer support in several different languages. One solution to this challenge is centralizing support with English speaking agents and having a translation layer in the middle to translate from the customer's language into the agent's (English) and vice versa. The ideal solution for this environment needs to consider the context of both sides which are in different languages, and also needs to be robust to the noisy input since the text here represents a higher degree of noise com-

[1] www.unbabel.com

[2] https://github.com/huggingface/neuralcoref

agent	src: Hi there! How can I help? tgt: Hallo! Wie kann ich helfen?
customer	src: Hey, ich muss mein Auto zum Mechaniker bringen und ich würde gerne Intelligent Auto Imports besuchen. tgt: Hey there, I need to take my car to mechanic and I would like to see Intelligent Auto imports.
agent	src: Sure! what type of car is it? tgt: Sicher! Was für ein Auto ist das?

Table 1: An example of a conversation between a customer and an agent.

	Customer		Agent	
	lines	words	lines	words
Training	6,216	41,492	7,629	70,193
Dev	862	5,805	1,040	9,569
Test	967	6,464	1,133	10,187

Table 2: Statistics of the English side of the training, dev, and test sets.

pared to the cases like news, biomedical, etc. In the first edition of this shared task we focused on this environment and asked the participants to translate the customer's utterances from German into English and the agent's from English into German.

Although participants were encouraged to submit both directions (i.e. modelling both speakers was desired), in this first round of the task, we emphasized on the agent side (English→German) and performed human evaluation in that direction exclusively. This decision is not entrenched and, thus, for future tasks we will aim at evaluating both translation directions. We decided to pursue this direction because the customer side (German→English) suffers from "translationese": English was the original source, and it was recently shown that translationese has a significant impact in evaluation both in automatic metrics (Freitag et al., 2020) and human evaluation (Läubli et al., 2020).

3.1 Data

The main data source for this shared task is ***BConTrasT***. As mentioned in §2, the translated conversations are sampled from the original Taskmaster-1 corpus, and in theory the other monolingual data could be leveraged by the participants either for back-translation or training in-domain language models. However, due to the high degree of sentence similarity within the Taskmaster-1 monolingual corpus, participants were not allowed to use this additional data to train their systems.

In addition to the provided in-domain training data, the participants were allowed to use all the training data provided by the News shared task organizers. Moreover, they were allowed to use existing pre-trained models, such as BERT (Devlin et al., 2018), Transformer-XL (Dai et al., 2019), Reformer (Kitaev et al., 2020), among others.

3.2 Baseline

To define our non-human baseline, we use Facebook's last year submissions to the document-level translation task for both directions (Ng et al., 2019) as the terms of comparison. Even though these models are not domain adapted for the Chat Translation task, we find them to have a reasonable quality for this domain. However, it is worth mentioning that we solely report the results of these models with the automatic metrics and we do not perform any type of direct assessment on these models.

4 Participants

Six participants submitted their systems to the Chat Translation shared task. Although the German→English direction (i.e. customer side) was optional, all participants submitted their systems for both directions. In total, 14 runs were submitted (although only primary submissions were considered for human evaluation). Table 3 summarizes the participants and their affiliations.

Team	Institution
NaverLabs	Naver Labs Europe
UEdinUppsala	Univ. of Edinburgh, Uppsala Univ.
IndTaoWang	Individual participant (Tao Wang)
Tencent	Tencent
UMaryland	University of Maryland
UJordan	Jordan U. of Science and Technology

Table 3: The participating teams and their affiliations.

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 Naver Labs

Naver Labs Europe (NLE) uses a document-level model trained on both the parallel and back-translated data. The authors developed a multi-domain system using the task-specific adapter layers and used it to participate in all the following tasks: chat translation, robustness, and biomedical. These systems are designed to translate both German and English text, or even mixed-language documents. Furthermore, in order to improve the robustness of these systems to noise, the authors applied the following pre-processing solutions: special handling of case with inline casing, a `copy` placeholder for rare characters, synthetic noise generation, and BPE dropout. Their primary submission is an ensemble of three instances of this model, which was used to decode the full bilingual dialogues at once using the entire dialogue's context. The first contrastive submission is a single model with these settings. The second submission is an ensemble of four sentence-level bidirectional models (one of them with masked language model pre-training). For more details see Bérard et al. (2020).

4.1.2 Universities of Edinburgh and Uppsala

The joint submissions of University of Edinburgh and Uppsala University are based on the transformer-big architecture (Vaswani et al., 2017) and rely on fine-tuning pre-existing systems from the WMT 2019 News Translation Task (experiment with both UEdin's submission based on Marian (Junczys-Dowmunt et al., 2018) and Facebook's submission based on Fairseq (Ott et al., 2019)). They are fine-tuned on pseudo-in-domain web crawled data and in-domain task data. The authors also experiment with (i) domain and speaker-level adaptation by automatically tagging the source and target sentences with domain and speaker tags respectively, and (ii) contextual NMT by exploiting the previous context, varying the type and number of previous utterances used. The final submission is an ensemble of four models trained with domain tags and using noisy-channel re-ranking. For more details see (Moghe et al., 2020).

4.1.3 Tao Wang (individual participant)

Individual participant Tao Wang uses a sentence-level system trained on all the WMT20 En-De parallel data. The author uses the Fairseq codebase to train a transformer-big model with the default settings of a base model. Then, the models are fine-tuned with the in-domain training set provided for the Chat Translation shared task.

4.1.4 Tencent

Tencent systems are based on self-attention networks including document-level multi-encoder and sentence-level Transformer. In order to get more in-domain data the authors use a multi-feature data selection method (e.g. FDA, n-gram LM, Transformer LM and BERT) to select data from news corpus. Furthermore, the systems have different fine-tuning strategies, ranging from sentence-level to document-level. Finally, these systems use large scale pre-trained language models including monolingual BERT (Devlin et al., 2018) and bilingual XLM (Lample and Conneau, 2019). For more details see (Wang et al., 2020).

4.1.5 University of Maryland

The University of Maryland systems are both sentence and document-level systems, with two distinct architectures for this task: (i) standard transformer pre-trained on WMT17 News and fine-tuned on the WMT20 Chat data, and (ii) modified transformer by including additional encoder to process one previous utterance in tandem with the current utterance, also pre-trained on WMT17 News and fine-tuned on a mix of WMT20 Chat data and a subset of WMT19 News data. The primary system is based on the first architecture while the second architecture is used for the two contrastive submissions. The contrastive submissions differ in the manner and timing in which training data was processed. For more details see (Bao et al., 2020).

4.1.6 Jordan University of Science and Technology

Mohammed et al. (2020) train separate models for the agent and customer sides after combining the training and development datasets for each side. They use bidirectional RNN (LSTM) with pre-trained BERT (Devlin et al., 2018) embeddings for each of the translation directions. In addition, the authors report using different parameters for training, resulting in different models which then are used for ensemble decoding. For more details see (Mohammed et al., 2020).

4.2 Submission Summary

The submissions for this year's shared task cover different approaches from simple sentence-level to more complex document-level models with extra encoders and decoders to summarize the context

(i.e. previous sentences), and from single direction to bi-directional translations (i.e. jointly modelling both En→De and De→En directions). Moreover, they report different approaches for training their systems ranging from fine-tuning the existing models and using embeddings of the large pre-trained models such as BERT (Devlin et al., 2018) to training the models from scratch.

Not only the submissions are different in their architectures, but they also differ in the data they use during the training. Some use all the available WMT parallel data in addition to the in-domain training data provided for the Chat task, and some apply data selection methods to get more in-domain data to leverage for training their systems.

5 Evaluation Procedures

For the first round of the Chat Translation shared task we follow the standard procedure of WMT shared tasks and evaluate both on automatic metrics and human evaluation with context. Even though automatic metrics provide a cheap mechanism to evaluate Machine Translation (MT) systems outputs, they do not tell the whole story for high-performing systems (Ma et al., 2019). For example, recent "sentence-level human parity" claims do not seem to hold when the context of the document is considered (Läubli et al., 2018), and metrics such as BLEU (Papineni et al., 2002) fail to correlate properly with human assessment (Callison-Burch et al., 2006). In this edition of the shared task, we aim for both automatic and manual evaluations.

5.1 Automatic Evaluation

For the automatic evaluation, we use both BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) metrics. For the former, we use SacreBLEU[3] (Post, 2018), while for TER we use v0.7.25[4] and report case-sensitive scores. The automatic metrics are used to measure the quality of the translations of both sides, i.e. customer and agent.

5.2 Human Evaluation

For the human evaluation we follow a similar procedure to last year's WMT News shared task (Barrault et al., 2019) but take into account the set of recommendations defined by Läubli et al. (2020).

[3]BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+tok.13a+version.1.4.13, BLEU+case.mixed+lang.de-en+numrefs.1+smooth.exp+tok.13a+version.1.4.13

[4]http://www.cs.umd.edu/~snover/tercom/

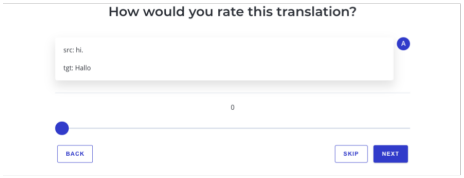
	Agent		Customer	
System	**BLEU↑**	**TER↓**	**BLEU↑**	**TER↓**
FAIR WMT'19	43.4	38.0	49.7	32.0
Primary				
NaverLabs	60.1	25.7	61.0	23.3
UEdinUppsala	60.2	25.4	**62.4**	**22.8**
IndTaoWang	59.7	26.0	61.3	23.5
Tencent	58.6	26.7	62.3	23.0
UniMaryland	56.7	28.2	49.4	32.0
UJordan	46.4	38.2	42.5	40.2
Contrastive				
NaverLabs-Sys1	58.8	26.8	59.4	24.6
NaverLabs-Sys2	**60.4**	**25.1**	61.6	23.1
UEdinUppsala-Sys1	60.2	25.3	61.8	22.8
UEdinUppsala-Sys2	59.8	25.4	61.5	23.8
Tencent-Sys1	53.6	30.6	54.0	28.8
Tencent-Sys2	58.6	26.6	61.9	23.2
UniMaryland-Sys1	55.6	28.3	49.4	32.0
UniMaryland-Sys2	56.4	28.1	49.4	32.0

Table 4: Automatic evaluation scores for the agent (En→De) and customer (De→En).

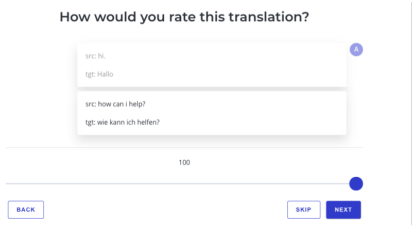
Specifically, we build *HITs* (following the Mechanical Turk's term *human intelligence task*) for the Segment Rating + Document Context (SR+DC) configuration with approximately 100 tasks similarly to WMT News, where both the source and target context is available to the evaluator when rating the actual source and target sentence for evaluation. We use an internal tool at Unbabel which provides the necessary visualization to evaluate a SR+DC configuration. Despite WMT News (Barrault et al., 2019) use Appraise (Federmann, 2012) for the human evaluation as it's tailored for document like text, the tool used for this task was built with chat evaluation in mind and outlines boundaries between each speaker. Figure 1 illustrates the tool used for evaluation.

Following Läubli et al. (2020) guidelines, we use trusted professional translators from the Unbabel community to evaluate the adequacy of the translation on a scale of 0 to 100. The guidelines to the translators were as simple as possible to avoid any type of bias, asking them to rate each sentence taking the context into account and penalizing when there is a context error, as they would for a non-contextual error.

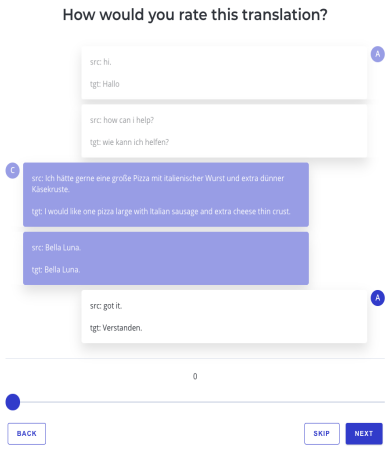
For the first edition of this shared task, we per-



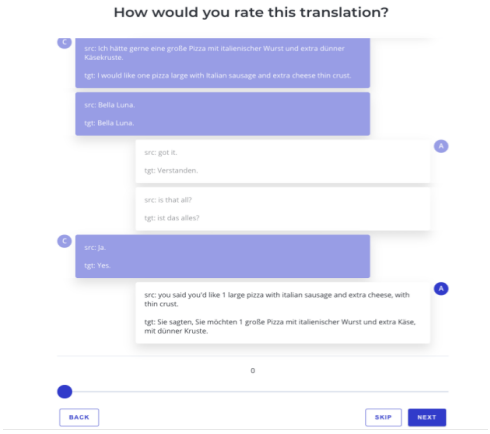
(a) First sentence of the conversation.



(b) Second sentence of the conversation.



(c) Third sentence of the agent in the conversation.



(d) Fifth sentence of the agent in the conversation.

Figure 1: Screenshots of a segment-rating with document-level context using the direct assessment tool. Multiple screenshots are presented to illustrate the iterative nature of the evaluation and how the agent and customer directions are presented as the conversation flows. Note that only the agent side is assessed and the scores are just illustrative.

formed human assessment on the agent side exclusively. Our decision is due to a limitation in the process of data creation, the *customer* direction is from professionally translated German (yet translated nonetheless) to the noisy original English (e.g. typos). Therefore, if we proceed with the evaluation as it stands we would induce two biases, 1) assessing a softer version of *translationese* as the source would be a translation, and 2) the noisy reference could bias the evaluators to rank the systems higher due to the noise and not quality. Both biases could be misleading and impacting their evaluations as professional translators are more sensitive to fine-grained phenomena (Läubli et al., 2020; Barrault et al., 2019). Moreover, in the proposed setting the impact of the noisy context for the agent is negligible for them to have a gist of the message; however there is an extra responsibility in translating the agent since the application of these systems in industry carries an additional factor: it has the company brand associated. Therefore, we preferred to focus more on evaluating the agent translations more rigorously than to spend resources in evaluating the customer.

5.2.1 Protocol for building HITs

We follow a hybrid between WMT News and Läubli et al. (2020) to build HITs. Specifically, as we resorted to professional translators there are fewer control tasks in every 100 HITs (i.e. 5% of the tasks being control tasks). To create a control task, we take inspiration from both the aforementioned resources and perform the following, assuming there is a vocabulary containing all the target words of the conversation: For the very short sentences containing one or two words we replace their words with some random words from the conversation's vocabulary. In the case of sentences with three words we replace the second and third words as before while keeping the first word. Finally, for longer sentences we preserve the first and last 10% of the words while randomly reordering the remaining 80% of the middle words. It is also worth mentioning that the corruption is only employed in the current sentence for evaluation and the context is preserved with no change.

When building the HIT bundle, among different options, we followed the same approach as WMT19 New's (Barrault et al., 2019) procedure

for SR+DC: in order to save time of our annotators, we built the HITs such that a sentence belonging to a given document is displayed and rated before the next sentence of the same document for the same participant MT system output. This is specially suited for our task as the conversations have larger contexts via numerous interactions. Similarly to WMT19 News (Barrault et al., 2019), we randomly picked documents from the pool of documents and for each participant retrieved their translations of that document. Next, we randomly picked documents from the pool until the sum of all their sentences was approximately 95 and added the remaining control tasks. For each document in the HIT, we sliced the translated conversation so that the order of the sentences was preserved when presented to the annotator for the SR+DC evaluation.

5.2.2 Evaluated Dialogs

Due to constraints with the annotators, we evaluated a subsample of the full test set. Therefore, we followed the procedure in § 5.2.1 with a budget constraint, where we specified the number of desired sentences and randomly sampled dialogues until the threshold is met (number of sentences). In the end, we evaluated 40% of the *agent* side, as noted in §5.2 we evaluated only this direction.

6 Discussion

The results of the automatic scores of both agent and customer side of all the submitted systems are reported in Table 4. Comparing these scores with our baselines (i.e. FAIR WMT'19 models) shows that in the agent side (En→De) there is a significant difference (i.e. between +3.0 to +17.0 BLEU scores) in the performance of the submitted systems and the baseline. However, comparing the differences between their TER scores reveals that there is a smaller gap between the systems, ranging from +0.2 to -12.9.

On the customer side we observe different behaviours and more diverse scores. In fact, the differences of the BLEU scores of the baseline and the submissions vary from -7.2 up to +12.7. This means that in a few cases our submitted systems fall behind the baseline by -7.2 BLEU scores. The TER scores show a similar behaviour and the differences of the scores of the submitted systems with the baseline varies from +8.2 (in the worst case) to -9.2 (in the case of best performing system). Given the fact that our references for this direction (i.e De→En) contain a higher degree of noise (eg. typos, wrong casings, etc) it is difficult to make a final and strong conclusion for this direction. We plan to investigate this aspect further.

	Agent	
System	**Avg.↑**	**Avg. z.↑**
Human	**91.4**	**0.319**
NaverLabs	88.2	0.165
UEdinUppsala	85.4	0.032
IndTaoWang	83.6	-0.049
UniMaryland	79.3	-0.235
Tencent	74.3	-0.474
UJordan	63.9	-0.966

Table 5: Human evaluation scores of the agent side.

Table 5 depicts the human evaluation scores (**Avg.**) and the normalized z-scores (**Avg. z**) of the agent side of the primary submissions. Human performance estimates are analogous to Barrault et al. (2019), evaluation of human-produced reference translations are denoted by "HUMAN" in all tables. There are three main clusters of scores, very high scores near human baseline levels (*NaverLabs*, *UEdinUppsala*, and *IndTaoWang*), significant scores (*UniMaryland* and *Tencent*), and lower scores (*UJordan*). Focusing on the high performing systems, we see that *NaverLabs* is the clear winner of the task, followed closely by *UEdinUppsala*, and *IndTaoWang*.

In addition to the overall DA scores of the submissions one might ask how they perform on the more detailed aspects such as sentences with different lengths or sentences containing pronouns. In order to address the first question, we analyzed the human scores for each system with respect to different intervals of lengths (i.e., different bins), namely 1-5 words, 6-10 words, 11-15 words, and, finally, 16+ words. To this end we can condition either (i) on the source sentence, or (ii) on the reference sentence, or (iii) on the generated translations of each system. Among these, we focused on (i) which provides more insights and is fairer comparison for all the systems.

Table 6 presents the human evaluation scores (**Avg.**) and the normalized z-scores (**Avg. z**) of the evaluated submissions in each length range. As we see, all the systems perform similarly in this range, all of them very close to the human reference. It is intersting to note that the submission of *UJordan* outperforms the human reference by +2.5

System	Source length range (words)							
	1-5		6-10		11-15		16+	
	Avg.	Avg. z.	Avg.	Avg. z.	Avg.	Avg. z.	Avg.	Avg. z.
Human	92.5	0.375	**92.5**	**0.367**	**90.0**	**0.254**	85.0	0.012
NaverLabs	92.5	0.375	86.9	0.103	88.3	0.170	**90.0**	**0.234**
UEdinUppsala	92.5	0.375	85.6	0.047	86.7	0.086	65.0	-0.936
IndTaoWang	92.5	0.360	83.1	-0.068	81.7	-0.146	75.0	-0.432
UniMaryland	90.0	0.249	79.4	-0.226	80.0	-0.210	55.0	-1.350
Tencent	85.0	0.042	71.9	-0.586	76.7	-0.378	65.0	-0.906
UJordan	**95.0**	**0.486**	71.3	-0.617	41.7	-2.012	10.0	-3.528

Table 6: Human evaluation scores of the agent side in each length range, based on the source sentences. The systems are ordered based on their general rankings.

	Agent	
System	**DA↑**	**z-score↑**
Human	**95.0**	**0.706**
NaverLabs	85.0	0.220
UEdinUppsala	85.0	0.220
Tencent	80.0	0.043
IndTaoWang	80.0	0.043
UniMaryland	80.0	0.043
UJordan	60.0	-0.861

Table 7: Human evaluation scores for the agent side when there is a pronoun *it* in the source sentence.

and +0.111 points on the average and normalized z-score, respectively. The differences increase by moving to the longer source sentences which is expected. The only unusual observation in these scores is the higher scores of the NaverLabs in the last range (i.e. sentences with 16+ words) in which it outperforms the human reference by +5.0 and +0.222 points on the average and normalized z-score, respectively. This can be due to the evaluators preferences, but still needs further analysis before making any final conclusion.

The English sentences containing pronouns is another aspect that we analyzed further and compared the performances of the submitted systems when there is a pronoun in the sentence. Specifically, we compute the scores for sentences which contain at least one instance of pronoun *it*. Table 7 shows the human scores and the normalized z-scores. As the results show, there is a big difference in the scores obtained by human translators and the submitted systems. In fact, it varies from -10.0 to -50.0 in the case of average score and from -0.486 to -1.567 for the normalized z-scores. Even though the number of tasks is not large, these preliminary results suggest current document-level systems still fall behind humans in challenging linguistic phenomena such as translating pronouns, and require further research for these phenomena.

Finally, we note that three of the submitted primary systems do not leverage the document-level context and use only the sentence-level information. Due to the data size and content proposed for the first edition of the Chat Translation shared task, this is to be expected as there is some level of repetition and similarity among different conversations. However, by looking at the results, we notice that approaches with document-level context seem to benefit from human evaluation when compared to the automatic metrics.

7 Conclusions

We presented the results of the first edition of the WMT20 Chat Translation shared task. For the purpose of this task, we created a bilingual English-German dialogue corpus, ***BConTrasT***, which is publicly available on the website of the task. It is based on the monolingual Taskmaster-1 corpus (Byrne et al., 2019) which was originally created in English. We translated around 18k of conversations of this corpus into German using the professional translators and used it as the in-domain corpus of the shared task.

This year we received 14 submissions from 6 different teams, all of them covering both directions (i.e. *customer* and *agent*). In addition to the automatic metrics (i.e. BLEU and TER) we performed an extensive Direct Assessment with document-level context using professional translators and used the results of these manual evalua-

tions to rank the participating systems. The previous sentences of each conversion provide the annotators with more context to have a more reliable assessment of the translations. Due to the constraints posed by our data, this year we were able to perform the manual evaluation only on the agent side (i.e. En→De). However, we aim at assessing both sides in the futures tasks.

Acknowledgments

We would also like to thank Mathieu Giquel and Ulisses Ferreira for all their help and support during the human evaluation phase, as well as Courtney Stankey for helping in coordinating with the evaluators. This work was supported by the P2020 programs MAIA (contract 045909) and Unbabel4EU (contract 042671), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UID/50008/2019.

References

Calvin Bao, Yow-Ting Shiue, Chujun Song, Jie S. Li, and Marine Carpuat. 2020. The university of maryland's submissions to the wmt20 chat translation task: Searching for more data to adapt discourse-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Alexandre Bérard, Vassilina Nikoulina, Ioan Calapodescu, and Jerin Philip. 2020. Naver labs europe's participation in the robustness, chat, and biomedical tasks at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*.

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.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*. Citeseer.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

António V Lopes, M Amin Farajian, Rachel Bawden, Michael Zhang, and André F T Martins. 2020. Document-level Neural MT: A Systematic Comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, André FT Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Nikita Moghe, Christian Hardmeier, and Rachel Bawden. 2020. The university of edinburgh-uppsala university's submission to the wmt 2020 chat translation task. In *Proceedings of the Fifth Conference on Machine Translation*.

Roweida Mohammed, Mahmoud Al-Ayyoub, and Malak Abdullah. 2020. Just system for wmt20 chat translation task. In *Proceedings of the Fifth Conference on 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)*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *In Proceedings of the Association for Machine Transaltion in the Americas (AMTA 2006*.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 113–123. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. 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 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020. Tencent AI Lab machine translation systems for the WMT20 chat translation task. In *Proceedings of the Fifth Conference on Machine Translation*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.