Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System

Mariana Neves[1*] **Antonio Jimeno Yepes**[2] **Aurélie Névéol**[3] **Rachel Bawden**[4]
Giorgio Maria Di Nunzio[11] **Roland Roller**[6] **Philippe Thomas**[6] **Federica Vezzani**[5]
Maika Vicente Navarro[7] **Lana Yeganova**[8] **Dina Wiemann**[9] **Cristian Grozea**[10]
[1]German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[2]RMIT University, Australia
[3]Université Paris-Saclay, CNRS, LISN, Orsay, France
[4]Inria, Paris, France
[5]Dept. of Linguistic and Literary Studies University of Padua, Italy
[6]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[7]Leica Biosystems, Australia
[8]NCBI/NLM/NIH, Bethesda, USA
[9]Novartis AG, Basel, Switzerland
[10]Fraunhofer Institute FOKUS, Berlin, Germany
[11]Dept. of Information Engineering, University of Padua, Italy

Abstract

We present an overview of the Biomedical Translation Task that was part of the Eighth Conference on Machine Translation (WMT23). The aim of the task was the automatic translation of biomedical abstracts from the PubMed database. It included twelve language directions, namely, French, Spanish, Portuguese, Italian, German, and Russian, from and into English. We received submissions from 18 systems and for all the test sets that we released. Our comparison system was based on ChatGPT 3.5 and performed very well in comparison to many of the submissions.

1 Introduction

We describe the eighth edition of the Biomedical Translation Task[1] that was part of the Eighth Conference on Machine Translation (WMT23). Similar to previous years, we released multiple test sets based on biomedical abstracts that we retrieved from the PubMed database.[2]

We addressed six languages pairs, namely German (de), Spanish (es), French (fr), Italian (it), Russian (ru), and Portuguese (pt), from and into English, as following:

- German into English (de2en) and English into German (en2de);
- Spanish into English (es2en) and English into Spanish (en2es);
- French into English (fr2en) and English into French (en2fr);
- Italian into English (it2en) and English into Italian (en2it);
- Russian into English (ru2en) and English into Russian (en2ru);
- Portuguese into English (pt2en) and English into Portuguese (en2pt).

Different from the previous editions of the shared task, we did not release test sets for Chinese–English or English–Chinese. Novel this year is that we relied on ChatGPT 3.5 to create a performance point of comparison (cf. Section 3), instead of our baseline systems from the previous years.

2 Test sets

We created the test sets following a similar procedure to previous years. We downloaded the set composed of daily update files from Pubmed[3] on

[*]The contribution of the authors are the following: MN prepared the MEDLINE test sets, performed test set validation, manual validation, and organized the shared task; AJY performed test set validation, manual validation, the automatic evaluation and co-organized the shared task; AN compiled information on participants' methods, performed test sets validation, manual validation and annotations of chatGPT outputs on the en2fr test set; RB, GMDN, RR, PT, FV, MVN, LY, DW performed test set validation and/or manual validation; and CG used OpenAI API to create the ChatGPT 3.5 point of comparison; All authors approved the final version of the manuscript. E-mail for contact: mariana.lara-neves@bfr.bund.de

[1]http://www2.statmt.org/wmt23/biomedical-translation-task.html

[2]https://pubmed.ncbi.nlm.nih.gov/

[3]https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 43–54
December 6–7, 2023.

April 26, 2023 and searched for articles that contained abstracts in both English and one of the six languages that we consider. We then randomly selected 100 bilingual abstracts for each of the language pairs.

For all language pairs, we split the sentences of the abstracts using SciSpacy (Neumann et al., 2019) and aligned them with the Geometric Mapping and Alignment (GMA) tool.[4] Native speakers of the languages manually checked the alignment quality in the Appraise tool (Federmann, 2018). In this evaluation, we classified the automatically aligned sentences into five categories:

1. "OK": both sentences contain the same information;
2. "Source>Target": the source sentence contains more information than the target one;
3. "Target>Source". the target sentence contains more information than the source one;
4. "Overlap": both source and target sentences have information not contained in the other one;
5. "No Alignment": the sentences refer to completely different contents, or one of hem is missing.

We present the results in Table 1. The highest alignment rates, i.e. the "OK" ones, were for Portuguese (at least 90%, both en2pt and pt2en), and the lowest ones for Russian (only 52% for en2ru). For the latter, we notice that the biggest difference with respect to the other language pairs is that many sentence pairs are not aligned, i.e. the "No Alignment" ones. The percentages for "Source>Target", "Target>Source", and "Overlap" are similar to the other language pairs. An analysis of these errors shows that they are due both to the sentence splitting and the alignment tool.

We released our test sets in two submission systems: (i) our Google form as announced on our shared task's web site; (ii) in OCELoT,[5] both in the General and in the Biomedical test sets.

3 Comparison system - ChatGPT 3.5

Instead of providing a baseline this year, we choose to provide translations from the ChatGPT 3.5 model through the OpenAI API. We refer to ChatGPT as a comparison system rather than a baseline, as it does not satisfy the usual criterion for a baseline as being a transparent, well-understood and reproducible model that provides a good (generally) lower bound against which to compare systems. Notably, the model is closed-sourced and trained on huge amounts of data, of which the details are not openly known.

ChatGPT 4 excels at many tasks (Chen et al., 2023; Jahan et al., 2023), including translation. Researchers from Tencent identified in a limited early evaluation done before the API was available ChatGPT 4 as a good translator (Jiao et al., 2023). Please note that we abstained from using the stronger ChatGPT 4 and used instead the faster but expectedly weaker ChatGPT 3.5. More precisely we used the model snapshot "gpt-3.5-turbo-0613", computed on June 13th 2023 but with the training data "up to Sept 2021"[6]. This reduces the risk of data contamination with respect to the abstracts used in our test sets, which were published in 2023.

The ChatGPT variants are large and trained on large quantities of data, but are generalist systems. Ideally, systems dedicated to translation or specialized in biomedical translation would be able to outperform them, or at least outperform the faster lower-quality version that we proposed here as a point of comparison. Otherwise, there are fewer reasons remaining for developing and using an alternative machine translation (MT) system: data privacy, self-hosting, usage in low-resources, non-connected systems.

We used the following prompt to perform the translations and to keep ChatGPT from producing any comments beyond the translation text itself: **"You are a helpful assistant specialised in biomedical translation. You will be provided with a sentence in {src}, and your task is to translate it into {trg}."** where {src} was the source language and {trg} was the target language (e.g. src = Italian and trg = English).

Using ChatGPT through the API proved to be more challenging than expected and seemed to act as a stress test for the API servers or for the cloudfare content distribution network proxy they use. For example we hit various intentional limitations, such as a rate limit of 90,000 tokens per minute. We then faced multiple other errors: read time out

[4] https://nlp.cs.nyu.edu/GMA/
[5] https://ocelot-wmt23.mteval.org/
[6] https://platform.openai.com/docs/models/gpt-3-5

Language	OK	Source>Target	Target>Source	Overlap	No Align.	Total
de2en	352 (82.2%)	20 (4.7%)	12 (2.8%)	9 (2.1%)	35 (8.2%)	428
en2de	471 (87.7%)	28 (5.2%)	9 (1.7%)	11 (2.0%)	18 (3.4%)	537
es2en	412 (89.5%)	16 (3.5%)	11 (2.4%)	-	21 (4.6%)	460
en2es	388 (88.4%)	21 (4.8%)	15 (3.4%)	6 (1.4%)	9 (2.0%)	439
fr2en	215 (85.3%)	17 (6.7%)	10 (4.0%)	7 (2.8%)	3 (1.2%)	252
en2fr	432 (83.7%)	78 (15.1%)	4 (0.8%)	-	2 (0.4%)	516
it2en	310 (73.4%)	46 (10.9%)	23 (5.5%)	6 (1.4%)	37 (8.8%)	422
en2it	298 (67.0%)	33 (7.4%)	29 (6.5%)	12 (2.7%)	73 (16.4%)	445
pt2en	385 (93.7%)	6 (1.4%)	7 (1.7%)	9 (2.2%)	4 (1.0%)	411
en2pt	450 (90.6%)	21 (4.2%)	12 (2.4%)	9 (1.8%)	5 (1.0%)	497
ru2en	233 (70.0%)	30 (9.0%)	16 (4.8%)	10 (3.0%)	44 (13.2%)	333
en2ru	221 (52.9%)	44 (10.5%)	23 (5.5%)	18 (4.3%)	112 (26.8%)	418

Table 1: Statistics (number of sentences and percentages) of the automatic alignment quality of the MEDLINE test sets.

in the object "HTTPSConnectionPool" with host api.openai.com, HTTP 502 (bad gateway), and "internal error". After writing our API calling code in an idempotent way, we were able to interrupt it whenever it was stuck and restart it whenever we stopped it or it stopped with an error. To this end, the script would skip over the existing translations and proceed with sending for translation, one by one, the rest of the entries not yet translated.

The overall experience remained positive, as building the ChatGPT 3.5 translations involved 674,470 tokens, resulting in a total API cost of only 1.15 USD. However, we have no information on the CO_2 impact of the computation, which should include the impact of inference for translations as well as a fraction of the impact of training the ChatGPT 3.5 model. Writing the scripts and executing them took less than three days. The execution itself was fast; as we reported here, at times we exceeded the API limit of 90,000 tokens per minute.

4 Teams and systems

After the release of the test sets, the teams had around two weeks to process the data and submit their translations. We collected submissions from the two systems (our Google form and OCELoT) belonging to 18 teams (or systems), as listed in Table 2. We allowed up to three runs for each team and language pair. From all submissions, we skipped only one translation from one team, namely the one for fr2en from UPCite-CLILLF, since it was in French (instead of English).

This year, the Google submission form also included questions on material and methods used by participants. The questions were identical to those used in 2022. The response rate was lower than in previous years (2020-2022) when the questionnaire was operated separately from the submission system and teams were asked to complete the survey after submission. In Ocelot submissions, participants were asked to submit a narrative description of their method. None of the teams reported the CO_2 impact of their participation in the task.

Many teams approached the task with transformer-based neural MT (NMT), relying on existing implementations. The use of prompting autoregressive models was also introduced this year. Table 3 presents details of the teams' methods.

5 Automatic evaluation

We present BLEU scores (Papineni et al., 2002) for the automatic evaluation in Tables 4 and 5. This includes translations received from both submission systems (Google Form and OCELoT).

For both en2de and de2en test sets, the submissions from HuaweiTSC, ZengHuiMT, GPT4-5shot, and PROMT teams obtained higher scores than our comparison system (ChatGPT) according to BLEU. The BLEU scores of the Lan-BridgeMT submissions (which use GPT3 and GPT4) came very close to those of ChatGPT for most language pairs, e.g., en2es, en2it, and were sometimes higher, e.g., fr2en, it2en, and ru2en. Most of the ONLINE system submissions also got higher BLEU scores than ChatGPT. However, it is worth bearing in mind the possibility that the ONLINE systems had previously seen our test sets in the large data on

Team ID	Institution	Biom. task	Publication
AIRC	Artificial Intelligence Research Center, Japan	-	(Rikters and Miwa, 2023)
GPT4-5shot	Microsoft	-	(Hendy et al., 2023)
GTCOM_Peter	Global Tone Communication, China	-	(Zong, 2023)
HuaweiTSC	Huawei Translation Service Center	Yes	(Wu et al., 2023)
Lan-BridgeMT	Lan-Bridge Communications, China	Yes	(Wu and Hu, 2023)
NLLB_Greedy	(unknown)	-	-
NLLB_MBR_BLEU	(unknown)	-	-
NRPU_FJWU	Fatima Jinnah Women University, Pakistan	Yes	(Firdous and Rauf, 2023)
ONLINE-A	(unknown)	-	-
ONLINE-B	(unknown)	-	-
ONLINE-G	(unknown)	-	-
ONLINE-M	(unknown)	-	-
ONLINE-W	(unknown)	-	-
ONLINE-Y	(unknown)	-	-
PROMT	PROMT LLC	-	(Molchanov and Kovalenko, 2023)
UPCite-CLILLF	Université Paris Cité, France	Yes	(Zhu et al., 2023)
ustc_ml_group	University of Science and Technology, China	Yes	-
ZengHuiMT	LanguageX, China	-	(Zeng, 2023)

Table 2: List of the participating teams and systems. The third column indicates the teams that directly participated on the Biomedical Translation Task.

Team ID	Language pair	MT method	Trained	Fine-Tuned	BT	LM
AIRC	en/de	Ensemble of Mega transformer models	Yes	No	Yes	Yes
GTCOM	en/de	Transformer model	-	-	-	multilingual models
HuaweiTSC	en/de	Transformer model	-	-	-	-
Lan-BridgeMT	en/de, en/es, en/fr, en/it, en/pt, en/ru	GPT prompting	No	No	No	GPT3, GPT4
NRPU_FJWU	en/fr	Fairseq NMT	No	Yes	No	No
PROMT	en/ru	Marian NMT	Yes	No	-	-
UPCite-CLILLF	en/fr	MBart-50	No	Yes	No	No
USTC	en/fr	Fairseq NMT	Yes	No	No	No
ZengHuiMT	en/de, en/ru	many-to-many encoder decoder transformer model	-	-	-	-

Table 3: Overview of methods used by participating teams. Information is self-reported through the Google/Ocelot submission form for each selected "best run". BT indicates if backtranslation is used and LM if language models were used.

which they were trained, or were used by the authors to assist the production of the abstracts used in the test sets. Although we use the ChatGPT model based on data prior to 2022, meaning that it could not be trained on the parallel abstracts used in the test sets, it is also possible that ChatGPT was used by authors to produce the abstracts that form part of the test set.

6 Manual evaluation

We carried out a manual validation of the quality of the translations for some language pairs using the "3-way ranking" task in the Appraise tool. It consists of a pairwise comparison with three text spans, for example for en2pt: (i) the source text in English, (ii) translation A in Portuguese, and (iii) translation B also in Portuguese. The text is either a sentence or the whole abstract, i.e., we carried out the validation for each sentence and then for the complete abstract.

The evaluator should choose one of the following four options: (i) A=B, i.e., both translations have similar quality; (ii) A>B, i.e., translation A is better than translation B; (iii) A<B, i.e., translation A is worse than translation B; and (iv) error flag in case one or both of the translations do not refer to the same source text.

For the language pairs that we considered, we randomly selected the abstracts until we had at least 100 sentences. We restricted the abstracts

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru
AIRC		0.3443					
GPT4-5shot		0.3881					0.3649
HuaweiTSC	run1*	*0.4369					
HuaweiTSC	run2	0.4345					
HuaweiTSC	run3	0.4422					
Lan-BridgeMT		0.3463	0.5098	0.5164	0.4640	0.4832	0.3361
NLLB_Greedy		0.3663					0.3461
NLLB_MBR_BLEU		0.3625					0.3504
ONLINE-A		0.4332					0.4125
ONLINE-B		0.4298					0.4648
ONLINE-G		0.4263					0.3939
ONLINE-M		0.3984					0.3827
ONLINE-W		0.4451					0.4083
ONLINE-Y		0.4075					0.4049
PROMT							0.3872
UPCite-CLILLF				0.2706			
ustc_ml_group	run1			0.4908			
ustc_ml_group	run2*			*0.4998			
ZengHuiMT		0.3883					0.3775
ChatGPT		0.3851	0.5097	0.5318	0.4607	0.5098	0.3513

Table 4: BLEU scores for "OK" aligned test sentences, from English. The submissions without a run number are the ones that were submitted to OCELoT. Primary runs are marked by *.

Teams	Runs	de2en	es2en	fr2en	it2en	pt2en	ru2en
AIRC		0.3714					
GPT4-5shot		0.4371					0.4774
GTCOM_Peter		0.4212					
HuaweiTSC		0.4771					
HuaweiTSC	run1*	*0.4778					
HuaweiTSC	run2	0.4776					
HuaweiTSC	run3	0.4853					
Lan-BridgeMT		0.4215	0.5769	0.4323	0.5272	0.5569	0.4750
NLLB_Greedy		0.4040					0.4386
NLLB_MBR_BLEU		0.3992					0.4437
NRPU_FJWU	run1*			*0.3350			
NRPU_FJWU(1)	run1			0.3082			
NRPU_FJWU	run2			0.2202			
NRPU_FJWU	run3			0.2395			
NRPU_FJWU(1)	run3			0.3350			
ONLINE-A		0.4606					0.5723
ONLINE-B		0.4662					0.4648
ONLINE-G		0.4364					0.5445
ONLINE-M		0.4465					0.4607
ONLINE-W		0.4759					0.4919
ONLINE-Y		0.4075					0.5089
PROMT							0.5156
UPCite-CLILLF							
ustc_ml_group	run1*			*0.4124			
ustc_ml_group	run2			0.3854			
ZengHuiMT		0.4316					0.5256
ChatGPT		0.4360	0.5827	0.4263	0.5067	0.5915	0.4417

Table 5: BLEU scores for "OK" aligned test sentences into English. The submissions without a run number are the ones that were submitted to OCELoT. Primary runs are marked by *.

to those in which the rate of well aligned (OK) sentences was at least 80%. We considered all pairwise combinations from the following translations: (i) the reference translation, as originally available in PubMed, (ii) translations from ChatGPT 3.5, and (iii) translations from systems that directly took part on the Biomedical Translation Task, and not only on the General Task (see Table 2).

We present the results in Tables 6 and 7. We compute a significance test (Wilcoxon test) when

comparing the systems (or reference translation) and we show in bold and with a star (⊛) those cases in which one system (or the reference translation) was better than the other one.

None of the teams could outperform the reference translation for all of the language pairs. Further, for all language pairs that we checked, the quality of the translations from ChatGPT was similar to the reference translation at the sentence level, i.e., there was no significant difference in the results. However, on the abstract level, the ChatGPT translations were found to be better than the reference translations for some language pairs, namely, en2ru and fr2en.

For some of the languages (e.g. en2de), the rankings from the automatic and manual translations appear consistent. The BLEU score from the HuaweiTSC team was much higher than the one from Lan-BrigdeMT (0.43 versus 0.35), and indeed, the quality of the translations from HuaweiTSC was better than the ones from Lan-BrigdeMT. There are however some differences in rankings. For example the manual rankings do not correspond exactly to the automatic rankings for ru2en, fr2en and en2it. Notably, ChatGPT appears to be penalised by BLEU and does better in the manual rankings.

6.1 Quality of the translations

We discuss below, for some language pairs, some of the mistakes that we observed during the manual validation of the submissions.

en2de Similarly to the last few years, the quality of the translations into German was very high. Overall, the individual translations were often similar and differed only in nuances, such as the order of the syntactic constituents. Some models seemed to favour compound nouns more often than others (e.g., Lammellentrennung vs Trennung der Lamellen). However, this usually had no impact on the translation quality. Some systems translated idioms, such as "window of opportunity", literally into German. Especially specialist terms were translated differently by the individual models and it was rather challenging to judge which of the translated terms has better quality (see Example 1).

(1) **en:** The most common surgical fixation options are cerclages and screws, ...
de$_1$: Die häufigsten chirurgischen Fixierungsoptionen sind Zerkel und Schrauben, ...
de$_2$: Die häufigsten chirurgischen Fixierungsmöglichkeiten sind Zuggurte und Schrauben, ...
de$_3$: Die häufigsten operativen Fixationsmöglichkeiten sind Cerclagen und Schrauben, ...

en2es As observed in the last few years, the overall quality of the translations into Spanish was very high. MT systems output was indistinguishable from human translations in many occasions for both systems evaluated: ChatGPT and Lan-Bridge.

The reference translation outperformed Lan-Bridge when evaluating sentences and abstracts. The reference translation was more consistent in the abstracts, had a higher fluency in the translation and a better choice of terminology than Lan-BridgeMT.

For example, "illness recurrence" was translated as "recurrencia' by Lan-Bridge, whereas the reference translation used a more appropriate term "recidiva". Another example in the translation of the term "coronary heart disease", that Lan-Bridge translates literaly as "enfermedad coronaria", while the reference translation uses the medical term "cardiopatía coronaria".

As mentioned, the reference translation was more fluent when compared to Lan-Bridge, oftetimes having a slightly better word order, better concordance subject/verb and using punctuation (commas and full stops) more fluently. Similarly, the reference translation slightly outperformed ChatGPT when comparing abstracts.

However the baseline translation was better than the reference translation at sentence level, this was due to a more overall fluent and consistent translation of abstracts observed in the reference translation when compared to ChatGPT. It must be noted that ChatGPT performed very well compared to the reference translation in most abstracts evaluated manually.

In the following example ChatGPT used the correct punctuation for numbers above 1,000 in Spanish and the reference translation used the incorrect punctuation and was penalized for this fact.

(2) **ChatGPT:** Se incluyeron un total de 22,148 pacientes de 40 estudios.
Reference: Se incluyó un total de 22.148 pacientes de 40 estudios.

When compared against each other, ChatGPT outperformed Lan-Bridge both at the abstract level

Lang. dir.	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
en2de	HuaweiTSC vs. reference	10	0	6	⊛ **4**	100	25	57	17
	HuaweiTSC vs. Lan-BridgeMT	10	⊛ **7**	2	1	100	⊛ **41**	54	5
	HuaweiTSC vs. ChatGPT	10	5	3	2	100	⊛ **29**	59	12
	reference vs. Lan-BridgeMT	10	6	3	1	100	⊛ **32**	54	3
	reference vs. ChatGPT	10	3	6	1	100	18	64	17
	Lan-BridgeMT vs. ChatGPT	10	0	3	⊛ **7**	100	10	61	⊛ **29**
en2es	ChatGPT vs. Lan-BridgeMT	13	4	6	3	107	21	71	15
	ChatGPT vs. reference	13	3	6	4	107	22	65	20
	Lan-BridgeMT vs. reference	13	1	6	6	107	14	69	24
en2fr	reference vs. Lan-BridgeMT	10	⊛ **9**	0	1	108	⊛ **80**	7	21
	reference vs. ChatGPT	10	7	1	2	108	⊛ **71**	5	32
	reference vs. UPCite-CLILLF	10	⊛ **10**	0	0	108	⊛ **107**	0	1
	reference vs. ustc_ml_group	10	⊛ **9**	0	1	108	⊛ **85**	1	21
	Lan-BridgeMT vs. ChatGPT	10	1	5	4	108	24	24	⊛ **60**
	Lan-BridgeMT vs. UPCite-CLILLF	10	⊛ **10**	0	0	108	⊛ **105**	3	0
	Lan-BridgeMT vs. ustc_ml_group	10	⊛ **8**	1	1	108	⊛ **54**	23	31
	ChatGPT vs. UPCite-CLILLF	10	⊛ **10**	0	0	108	⊛ **103**	3	2
	ChatGPT vs. ustc_ml_group	10	⊛ **9**	1	0	108	⊛ **73**	14	20
	UPCite-CLILLF vs. ustc_ml_group	10	0	0	⊛ **10**	108	7	4	⊛ **97**
en2it	Lan-BridgeMT vs. ChatGPT	15	2	1	⊛ **12**	92	16	29	⊛ **47**
	Lan-BridgeMT vs. reference	15	4	1	10	92	25	31	36
	ChatGPT vs. reference	15	9	1	5	92	24	31	37
en2pt	reference vs. Lan-BridgeMT	11	⊛ **5**	6	0	105	35	45	25
	reference vs. ChatGPT	11	4	4	3	105	25	48	32
	Lan-BridgeMT vs. ChatGPT	11	1	5	5	105	18	62	25
en2ru	reference vs. ChatGPT	13	4	3	6	94	8	60	⊛ **25**
	reference vs. Lan-BridgeMT	13	5	4	4	94	22	47	25
	ChatGPT vs. Lan-BridgeMT	13	7	3	3	94	20	64	10

Table 6: Pairwise manual evaluation results for the MEDLINE abstracts test set (from English). We show in bold (and with ⊛) the values which were statistically significant (Wilcoxon test).

Lang. dir.	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
fr2en	NRPU_FJWU vs. reference	19	1	0	⊛ **18**	108	18	6	⊛ **83**
	NRPU_FJWU vs. ustc_ml_group	19	1	1	⊛ **17**	108	19	11	⊛ **78**
	NRPU_FJWU vs. ChatGPT	19	0	0	⊛ **19**	108	3	8	⊛ **97**
	NRPU_FJWU vs. Lan-BridgeMT	19	3	3	⊛ **13**	108	25	11	⊛ **72**
	reference vs. ustc_ml_group	19	⊛ **12**	4	3	108	47	26	34
	reference vs. ChatGPT	19	5	7	7	108	30	26	⊛ **51**
	reference vs. Lan-BridgeMT	19	⊛ **15**	1	3	108	⊛ **60**	19	28
	ustc_ml_group vs. ChatGPT	19	0	1	⊛ **18**	108	13	39	⊛ **56**
	ustc_ml_group vs. Lan-BridgeMT	19	9	3	7	108	45	26	37
	ChatGPT vs. Lan-BridgeMT	19	⊛ **19**	0	0	108	⊛ **69**	30	9
ru2en	ChatGPT vs. reference	13	3	6	4	75	20	41	14
	ChatGPT vs. Lan-BridgeMT	13	⊛ **7**	5	1	75	⊛ **34**	37	4
	reference vs. Lan-BridgeMT	13	⊛ **9**	4	0	75	⊛ **42**	27	6

Table 7: Pairwise manual evaluation results for the MEDLINE abstracts test set (into English). We show in bold (and with ⊛) the values which were statistically significant (Wilcoxon test).

and at the sentence level. As with the reference translation, ChatGPT was more fluent, had a better choice of terminology (domain specific terms) and was more consistent overall at abstract level.

The ChatGPT translation was more fluent in the following example with a better usage of wording. Lan-bridge followed the English source text more closely which made the output less idiomatic.

(3) **ChatGPT:** IO redujo los niveles de glucosa en sangre, restableció el peso corporal y mejoró

la sensibilidad a la insulina, así como la tolerancia a la insulina y a la glucosa en ratones diabéticos.
Lan-bridge: IO redujo los niveles de glucosa en sangre, restableció el peso corporal y mejoró la sensibilidad a la insulina junto con la tolerancia a la insulina y la tolerancia a la glucosa en ratones diabéticos.

While issues are still being observed by the MT systems evaluated manually this year, these are no longer major translation issues as in past years. The issues observed this year for the translations from English to Spanish were minor issues that affect the overall final quality, but can be remediated by editing the MT output to provide better terminology, specially domain specific, more fluent sentences and a better overall consistency in the translation (specially for abstracts).

en2fr Translation quality was somewhat uneven this year. While some translations were very high quality and often similar or identical to reference translations, others exhibited serious issues including inserting erroneous information (see Example 4) or conveying meaning drastically different (see Example 5) or opposite to the original sentence (see Example 6). This type of error can have a severe impact when it results in incorrect medical information (Example 6) or incorrect description of a social group (see Example 5).

(4) **en:** Analysis (. . .) showed that. . .
fr$_1$: L'analyse (. . .) a montré que. . .
fr$_2$: * L'analyse (. . .) a montré que. . . (Traduit par Docteur Serge Messier)

(5) **en:** The criminalization of Black people
fr$_1$: La criminalisation des Noirs
fr$_2$: * La criminalisation des personnes blanches

(6) **en:** blood potassium level ≥ 6.5 mmol/L
fr$_1$: taux de potassium sanguin supérieur à 6,5 mmol/L
fr$_2$: * taux sanguin de potassium inférieur à 6,5 mmol/L

The translation of numerical values was also unreliable: example 5 illustrates the adequate translation of *6.5 mmol/L* into *6,5 mmol/L*, however in another abstract the study population of *52 dogs* was erroneously translated by *54 chiens*.

Issues remain with acronym translation where acronyms are often kept verbatim upon definition (e.g., *developmental disabilities (DD)* translated as *troubles du développement (DD)* instead of the reference translation *troubles du développement (TD)* although consistency seems improved: acronyms, albeit erroneous, are often used throughout a text.

The comparison of translations exhibiting different types of issues also remains difficult. In example 7, although *enquête* is a better translation for *survey* in the context, translation **fr$_1$** was preferred to **fr$_2$** because of the correct translation for *asking about*, which was central to the sentence.

(7) **en:** A survey asking about training
fr$_1$: Un sondage demandant des informations sur la formation
fr$_2$: * Une enquête demandant une formation

Overall, the one-to-one comparisons seemed quite consistent in ranking the systems and reference, and suggest that perhaps the most serious issues identified were concentrated in a few systems.

In addition to the manual evaluation through appraise, a complementary assessment of ChatGPT outputs was conducted, with a focus on *Acronyms* and *Lab Values*, which had been studied in our clinical case descriptions last year. We found that overall, 39 out of 50 test documents contained acronyms and only 3 contained lab values. The low frequency of lab values in the test set suggests that this particular source of translation difficulty for automatic system is not present in random scientifc abstracts. Furthermore, we cannot draw conclusions on the performance of ChatGPT on lab value translations. *Acronym* translations were considered correct when the ChatGPT translation was identical to the reference translation or consisted of an attested acronym use in similar context. Correct acronym translations (74%) included frequent acronyms such as CI (confidence interval), OR (odds ratio) or MRI (magnetic resonance imaging). In other cases, acronyms were either untranslated (16%) or erroneous (10%). These cases included acronyms for terms that were unfrequent or ad-hoc to the documents - albeit often a major topic. It should be noted that they were a source of inconsistent acronym translations in 14 documents - 36% of test documents with acronyms.

fr2en Translation quality was good overall and sometimes indistinguishable from reference translations. Aside from a problem with certain words being dropped at the beginning of translations, sometimes mid-word (quite possibly due to a bug by one or several of the systems), the errors made were similar to previous years.

Term and acronym translation (see Example 8) remained a serious problem and one that was highly influential in reranking decisions, i.e. more so than other errors such as those involving grammar, style or naturalness. In addition to acronym translation errors, we also observed that acronym placement was not always coherent (e.g. an acronym not being defined at the first instance and used consistently afterwards), but in practice this did not influence reranking decisions because of the presence of more serious errors.[7]

(8) **fr:** La migraine est la maladie neurologique la plus fréquemment rencontrée...
en$_1$: Migraine is the most common neurological disorder...
en$_2$: *Mimine is the most frequently encountered neurological disease...

The translation of non-domain-specific terms also posed problem, either those that were ambiguous in context (Example 9), including pronoun translation (for example *sa/son* 'his/her/its/their' being translated as *its* rather than 'his/their' or involving some degree of polarity (Example 10). On a similar note, the omission of words, mainly adjectives and adverbs (e.g. *relativement* 'relatively' and *souvent* 'often') sometimes made the difference between two translations, as did missing final punctuation (when no other errors were present).

(9) **fr:** ...les traitements oraux anciens...
en$_1$: ...older oral treatments...
en$_2$: *...ancient oral treatments...

(10) **fr:** ...un profil d'effets indésirables peu favorable
en$_1$: ...an unfavorable adverse effect profile
en$_2$: *...a slightly favorable side effect profile.

Finally, as in previous years, not all reference translations of were entirely faithful to the French source abstract (paraphrasing, missing or added information). This resulted in some cases in the reference translation being ranked below a system output, including imperfect outputs. Caution should therefore be taken when drawing conclusions about translation quality concerning humans, since intentional paraphrasing by the authors resulted in good abstracts but inferior in terms of our manual evaluation criteria. This partly explains why ChatGPT is "better" than the reference translations for this language pair.

en2it The quality of the translation was on average higher than the previous years. Most of the sentences compared was almost identical and fluent in terms of the quality of language. From a terminological viewpoint, it is possible to identify some inaccuracies in the choice of translating terms in the target language. For example, in *tumour recurrence*, the correct translation of *recurrence* is *recidiva* instead of *ricorrenza*.

Another frequent mistake, which is also a frequent mistake for language learners, is the translation of *hair* in sentences like "hair cortisol concentration (HCC) in healthy and ill cows". In these cases, *hair* must be considered as the hair of animals of body parts, therefore *peli*, and not scalp hair, in Italian *capelli*.

In some cases, there were better choices made by the reference system. For example, in the case of the phrase "[the author] is an initiate into the topic", ChatGPT used *iniziato* to translate *initiate* while a better equivalent would be in this case *novizio* as proposed by the reference system.

Finally, from a syntactic point of view, the results were very similar and only in a few cases we could find a construction that sounded odd or not easy to read. For example, the sentence "Flowmetry data always showed a more or less sudden disappearance of vasomotion." was translated by the reference system with *I dati della flussometria hanno sempre mostrato una più o meno improvvisa scomparsa della vasomotricità* while it would be more appropriate the translation of provided by the baselinte *I dati di flussometria mostravano sempre una scomparsa più o meno improvvisa della vasomozione.*

en2pt The results show that many translations, either from the referenc, ChatGPT, or from the Lan-BridgeMT team, were as good as the reference translation for many sentences (cf. Table 6,

[7] This could be something to look out for in future years when evaluating whole abstracts, when the translation quality allows such fine-grained observations.

"Sentences"). However, there were many cases on which we decide that one passage was better than the other, we discuss some of these differences here.

The most serious mistake that we found was the translation of "back pain" into "pressão arterial" (blood pressure), probably because both of them have the same acronym in English, i.e., "BP".

(11) **en:** The high incidence and worsening of BP ...
pt$_1$: A alta incidência e agravamento do PC ...
pt$_2$: A alta incidência e o agravamento da pressão arterial ...

Similar to previous years, we still found cases in which the English (or simply a wrong) acronym was used (cf. exmple below). Some similar errors might only be noticed when checking the complete text (abstract), and not only single sentences, such as when the translation includes an acronym that was not defined previously.

(12) **en:** ... Creutzfeldt-Jakob disease (CJD) ...
pt$_1$: ... doença de Creutzfeldt-Jakob (DCJ) ...
pt$_2$: ... doença de Creutzfeldt-Jakob (CJD) ...

In some cases, even though both passages were correct, we found that the translation was better due to the use or more medical concepts.

(13) **en:** ... headache attributed to ischemic stroke ...
pt$_1$: A cefaleia atribuída ao acidente vascular cerebral isquêmico ...
pt$_2$: ... a dor de cabeça atribuída ao derrame isquêmico ...

Sometimes the translation included terms that were not suitable, even thought the meaning was close to the source, and it the might have been understood by many readers.

(14) **en:** ... which were analyzed fully and individually.
pt$_1$: ... que foram analisados na íntegra individualmente.
pt$_2$: ... que foram analisados de forma completa e individual.

We chose translation which better describe the facts, depending of the use active or passive voice. Further, in case of passive voice, we preferred caes in which the subjective is closer to the verb, or even before it. We find that it improves the readability.

(15) **en:** The patients underwent magnetic resonance imaging.
pt$_1$: Os pacientes realizaram ressonância magnética. (active voice)
pt$_2$: Os pacientes foram submetidos a ressonância magnética. (passive voice)

(16) **en:** Twelve articles were included in the analysis.
pt$_1$: Foram incluídos na análise 12 artigos.
pt$_2$: Doze artigos foram incluídos na análise.

ru2en While the quality of ru2en translations continues to impress, one recurrent issue centers around the proper handling of abbreviations and acronyms. Often, an acronym is introduced early in the abstract, and holds a clear, defined meaning. Yet, as the text progresses, these acronyms are frequently mishandled by translation systems, failing to link them to their previously established acronym, and frequently transliterating an acronym created in Russian text. This issue manifests itself nearly every time an acronym appears, which makes translations of abstracts that include acronyms not consistently reliable.

For example, the term Ischemic Stroke is introduced in the abstract and abbreviated to "ИИ" which corresponds to the Russian term "ишемического инсульта". One of the reference translations correctly uses the acronym IS to refer to Ischemic Stroke, while the other comes up with an unrelated abbreviation AI.

(17) **ru:** В исследование включили 120 пациентов (57 женщин и 63 мужчины, средний возраст 58,4±6,4 года) в позднем восстановительном периоде ИИ.
en$_1$: The study included 120 patients in the late recovery period of IS, 57 women and 63 men, average age 58.4±6.4 years.
en$_2$: The study included 120 patients (57 women and 63 men, median age 58.4±6.4 years) in the late recovery period of AI.

7 Conclusions

We presented the finding of the edition of the WMT Biomedical Translation Task. We received submission from 18 systems and compared them to translations from ChatGPT 3.5.

In the automatic evaluation, some systems were scored higher than BLEU according to the comparison system (ChatGPT 3.5). In the manual evaluation, none of the systems were systematically better

than the reference translation for all of the language pairs that we evaluated. However, in a couple of cases, namely, for fr2en and en2ru, the translations from ChatGPT were preferred over the reference translations. We presented a details discussion of the errors that we found during the manual evaluation.

Limitations

Our test sets comprise 50 abstracts per language pair/directions. Further, due to the time consuming, difficulty of the task, and number of submissions, the manual evaluation was only carried out for a small sample. However, since our task has been running for eight years, the cumulative number of test sets is satisfactory for testing purposes, and maybe even for few-shot training approaches.

We did not carry out manual evaluation for some of the language pairs (directions), e.g., it2en, for which we do not have experts who are native speakers in the target language and have a very good knowledge in the source language. However, we always release the test sets and the submission files from the participants, with which anyone can carry out further experiments or manual evaluations.

Ethics Statement

Our test sets were derived from PubMed, a database of biomedical citations. These publications are often used in many areas of the medicine, including decision about diagnostic and treatment of patients. Automatic translation in this domain should be used as part of a larger framework that should include human experts for the interpretation of the translations and, if necessary, correct and adapt the text accordingly.

Acknowledgements

Rachel Bawden's participation was funded by her chair position in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by the Emergence project, DadaNMT, funded by Sorbonne Université.

References

Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. 2023. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics*, 39(9):btad557.

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

Sheema Firdous and Sadaf Abdul Rauf. 2023. Biomedical Parallel Sentence Retrieval using Large Language Models. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good are GPT Models at Machine Translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Alexander Molchanov and Vladislav Kovalenko. 2023. PROMT Systems for WMT23 Shared General Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. 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.

Matīss Rikters and Makoto Miwa. 2023. AIST AIRC Submissions to the WMT23 Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. The Path to Continuous Domain Adaptation Improvements by HW-TSC for the WMT23 Biomedical Translation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hui Zeng. 2023. Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Lichao Zhu, Maria Zimina-Poirot, Maud Bénard, Behnoosh Namdar, Nicolas Ballier, Guillaume Wisniewski, and Jean-Baptiste Yunès. 2023. Training data filtering and fine-tuning strategies - discoveries of UPCite-CLILLF Team's participation in WMT 23 Biomedical Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hao Zong. 2023. GTCOM Neural Machine Translation Systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.