Findings of the WMT 2022 Biomedical Translation Shared Task: Monolingual Clinical Case Reports

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, et al.

Findings of the WMT 2022 Biomedical Translation Shared Task: Monolingual Clinical Case Reports

Mariana Neves[1] ***Antonio Jimeno Yepes**[2] **Amy Siu**[3]
Roland Roller[4] **Philippe Thomas**[4] **Maika Vicente Navarro**[5]
Lana Yeganova[6] **Dina Wiemann**[7] **Giorgio Maria Di Nunzio**[8]
Federica Vezzani[8] **Christel Gerardin**[9] **Rachel Bawden**[10]
Darryl Johan Estrada[11] **Salvador Lima-López**[11] **Eulàlia Farré-Maduell**[11]
Martin Krallinger[11] **Cristian Grozea**[12] **Aurélie Névéol**[13]
[1]German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[2]RMIT University, Australia
[3]Berliner Hochschule für Technik, Germany
[4]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[5]Leica Biosystems, Australia
[6]NCBI/NLM/NIH, Bethesda, USA
[7]Novartis AG, Basel, Switzerland
[8]Dept. of Linguistic and Literary Studies University of Padua, Italy
[9]Sorbonne Université, Inserm, IPLESP, Paris, France
[10]Inria, Paris, France
[11]Barcelona Supercomputing Center, Spain
[12]Fraunhofer Institute FOKUS, Berlin, Germany
[13]Université Paris-Saclay, CNRS, LISN, Orsay, France

Abstract

In the seventh edition of the WMT Biomedical Task, we addressed a total of seven language pairs, namely English/German, English/French, English/Spanish, English/Portuguese, English/Chinese, English/Russian, English/Italian. This year's test sets covered three types of biomedical text genre. In addition to scientific abstracts and terminology items used in previous editions, we released test sets of clinical cases. The evaluation of clinical cases translations were given special attention by involving clinicians in the preparation of reference translations and manual evaluation. For the main MEDLINE test sets, we received a total of 609 submissions from 37 teams. For the ClinSpEn sub-task, we had the participation of five teams.

1 Introduction

This is the seventh edition of the biomedical translation task offered under the umbrella of the Conference on Machine Translation (WMT22).[1] This shared task builds on the six previous editions of the biomedical translation task (Bojar et al., 2016; Jimeno Yepes et al., 2017; Neves et al., 2018; Bawden et al., 2019, 2020; Yeganova et al., 2021). Similar to previous years, we addressed seven language pairs, in both directions, namely: German/English (de2en and en2de), Spanish/English (es2en and en2es), French/English (fr2en and en2fr), Italian/English (it2en and en2it), Portuguese/English (pt2en and en2pt), Russian/English (ru2en and en2ru), and Chinese/English (zh2en and en2zh).

In the biomedical translation task this year, participants were asked to translate shared test sets (described in Section 2) comprising documents belonging to three different text genres: scientific abstracts, clinical cases and terminology items. In total, seven language pairs (14 translation directions) were included this year, with both low-resource and high-resource pairs. For each language direction, we provide baseline systems relying on pre-trained

*The contribution of the authors are the following: MN prepared the MEDLINE test sets, performed test set validation, manual validation, and organized the task; AJY performed test sets validation, manual validation and the automatic evaluation; RR, PT, MVN, LY, DW. GMDN, FV performed test sets validation and manual validation; CGn created reference translation and performed manual validation; RB performed manual validation; DJE, SLL, EFM, MK organized the ClinSpEn sub-task; CGa created the baselines; and AN collected information on participants' methods, performed test sets validation, manual validation and created reference translation. All authors approved the final version of the manuscript. E-mail for contact: mariana.lara-neves@bfr.bund.de

[1] https://www.statmt.org/wmt22/biomedical-translation-task.html

neural translation models (described in Section 3). In order to help gain insight into the system performance, we collected information on the specific material and methods used in the systems from the participants (Section 4). System outputs for each task were evaluated both automatically and manually (as described in Section 5 and 6, respectively). One particular growth direction that we explored this year was the inclusion of full clinical case descriptions. A small set of five clinical cases in English were included in the MEDLINE test sets from English and a larger clinical corpus corpus was also included in the ClinSpEn track. We also involved clinicians in the preparation of gold standard translations and manual evaluation of clinical cases for en2fr and en2es.

Two types of submissions were received for the MEDLINE test sets: those submitted using (i) our submission system (hereafter called BioWMT), as in previous years, and (ii) the OCELoT submission system,[2] which was also used in the WMT general task.

In addition, an independent subtask was held as part of the Shared Task: ClinSpEn.[3] ClinSpEn focuses on the automatic translation of clinical content in both English and Spanish. Three subtracks are proposed based on different possible use cases: clinical case reports, clinical terminology obtained from literature and Electronic Health Records (EHR) and ontology concepts. Unlike the rest of the tasks, ClinSpEn's evaluation was done through CodaLab[4] and new submissions can still be made.

2 Test sets

In this section we describe the various test sets that we released for this year's edition of the WMT Biomedical task.

2.1 MEDLINE test sets

The MEDLINE test sets consisted of abstracts and case reports from the MEDLINE database. We aimed to retrieve 50 articles for each language direction. For the directions into English, the test sets consisted only of parallel abstracts. For the directions from English, we manually selected five clinical case reports, which were only available in English and which were the same across all language pairs. We completed each of these test sets with parallel abstracts. Table 1 summarizes the MEDLINE test sets as released in our submission system and in OCELoT. The only difference between the test sets released in our submission system and in OCELoT was that the latter contained aligned sentences, as provided by the automatic alignment. We describe the construction of the parallel abstracts and clinical case reports below.

2.1.1 Parallel abstracts

For the parallel abstracts, we downloaded the MEDLINE database[5] around the end of February and selected parallel abstracts for each language pair. We targeted publications whose PMID (PubMed identifier) was not included in any of our previous test sets and training data. We processed the abstracts using the same tools for sentence splitting and sentence alignment as in previous years (Yeganova et al., 2021). We manually checked the quality of the alignment using the Appraise tool (Federmann, 2010) and present results in Table 2.

2.1.2 Clinical case reports

For test sets *from English*, we decided to select clinical case presentations in order to include documents that would be closer in genre to clinical narratives found in patient records. Five clinical cases[6] were selected from publications of the *Journal of Medical Case Reports* (an open access publication) according to the following criteria:

- Reports a case related to oncology (based on the expertise of clinicians that agreed to contribute to the evaluation);
- Reports containing specific values such as lab results;
- Reports containing a limited amount of references to images and tables (to maximize resemblance with EHR narrative);

Both the abstract of the article and the full case presentation were included in the test set.

A gold standard translation of the clinical cases (both abstract and full case presentations) was created for French. We used the free version of

[2] https://github.com/AppraiseDev/OCELoT

[3] https://temu.bsc.es/clinspen

[4] https://codalab.lisn.upsaclay.fr/competitions/6696

[5] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

[6] PMIDs: 19144122, 21838907, 35303936, 35313981, 35144678

Pairs	Documents			Sentences (WMTBio)		Sentences (OCELoT)	
	Mono.	Parallel	Total	Mono.	Parallel	Mono.	Parallel
de2en	-	50	50	-	434/453	-	419
en2de	5	45	50	210/-	462/467	210	435
es2en	-	50	50	-	459/461	-	436
en2es	5	45	50	210/-	397/404	210	377
fr2en	-	50	50	-	319/325	-	308
en2fr	5	45	50	210/-	608/609	210	590
it2en	-	43	43	-	457/461	-	427
en2it	5	39	44	210/-	372/364	210	327
pt2en	-	50	50	-	459/478	-	454
en2pt	5	45	50	210/-	465/454	210	443
ru2en	-	50	50	-	408/398	-	351
en2ru	5	45	50	210/-	526/545	210	453
zh2en	-	48	48	-	281/409	-	277
en2zh	5	45	50	210/-	424/362	210	359

Table 1: Number of documents and sentences in the MEDLINE test sets. For the Ocelot test sets, the test sets have the same number of sentences for both languages in a pair.

Language	OK	Source>Target	Target>Source	Overlap	No Align.	Total
de2en	358 (85.2%)	26 (6.2%)	14 (3.3%)	7 (1.7%)	15 (3.6%)	420
en2de	383 (87.0%)	28 (6.4%)	13 (3.0%)	4 (0.9%)	12 (2.7%)	440
es2en	367 (83.4%)	32 (7.3%)	11 (2.5%)	11 (2.5%)	19 (4.3%)	440
en2es	350 (90.9%)	11 (2.9%)	14 (3.6%)	2 (0.5%)	8 (2.1%)	385
fr2en	253 (84.7%)	21 (7.0%)	6 (2.0%)	1 (0.3%)	18 (6.0%)	299
fr2en §	288 (93.6%)	5 (1.6%)	5 (1.6%)	2 (0.6%)	8 (2.6%)	308
en2fr	450 (86.8%)	64 (12.4%)	1 (0.2%)	-	3 (0.6%)	518
en2fr §	590 (97.8%)	13 (2.2%)	-	-	-	603
it2en	340 (79.0%)	44 (10.2%)	19 (4.4%)	14 (3.2%)	14 (3.2%)	431
en2it	261 (75.9%)	21 (6.1%)	16 (4.6%)	4 (1.2%)	42 (12.2%)	344
pt2en	426 (93.8%)	17 (3.7%)	8 (1.8%)	3 (0.7%)	-	454
en2pt	365 (82.2%)	36 (8.2%)	14 (3.1%)	7 (1.6%)	22 (4.9%)	444
ru2en	226 (64.4%)	25 (7.1%)	17 (4.8%)	7 (2.0%)	76 (21.7%)	351
en2ru	281 (61.2%)	32 (7.0%)	30 (6.5%)	25 (5.5%)	91 (19.8%)	459
zh2en	264 (94.0%)	4 (1.4%)	8 (2.8%)	-	5 (1.8%)	281
en2zh	346 (95.9%)	3 (0.8%)	5 (1.4%)	-	7 (1.9%)	361

Table 2: Statistics (number of sentences and percentages) of the quality of the automatic alignment for the MEDLINE test sets. § Results after manual correction of sentence segmentation and/or alignment.

DeepL[7] followed by two rounds of post-edition: first, a native French speaker with formal translation training and knowledge of clinical text (AN) post-edited the machine translation (MT) focusing on linguistic quality and fluidity of the translation; second, a clinician (CG) post-edited the revised text focusing on clinical correctness and adequacy of the text with the French clinical narrative genre. In this second step, special attention was given to values such as lab results, which can be expressed using different units in English vs. French. The goal was to produce a translation that would convey properly processed information for direct use by a clinician. We computed BLEU scores between the original machine translated text and successive rounds of post-edition. BLEU between MT and the final gold standard translation was 38 for abstracts and 42 for full texts, while BLEU between the translator post-edited text and final gold standard translation was 63 for abstracts and 85 for full texts.

[7] http://www.DeepL.com/Translator

2.2 ClinSpEn test sets

For each of the ClinSpEn sub-tracks, a gold standard dataset was prepared with human translations created by domain experts. Additionally, a big collection of monolingual background data was provided for each subtrack so that participants could test the scalability of their systems or use them for other purposes.

Sub-track 1: Clinical case reports. This subtrack deals with the translation of clinical case reports. Clinical cases are a text genre where a patient's current condition, medical history, clinical presentation, examinations, treatment and diagnosis are described. They can be pretty similar to EHR both in form and content. However, unlike EHR, clinical cases are often free of privacy-related issues. This means that they can be used as substitute to train NLP systems for the clinical domain.

The gold standard dataset's clinical cases were carefully selected to cover a wide range of aspects related to COVID-19: different types of patients (children, adults, elderly and pregnant people, babies), different comorbidities (cancer, mental health issues, immunosuppressed patients) and symptomatology (mild and severe presentations, dermatologic, immunologic and psychiatric manifestations, thrombosis, etc.). The reports were translated from English to Spanish by a professional medical translator in a first step and revised by a clinical expert in a second step. The background set includes around 3,800 clinical case reports in English extracted from PubMed Central.

The dataset includes a total of 202 COVID-19 clinical case reports (50 for the dev set, 152 for the test set) and the direction of this sub-track is en2es.

Sub-track 2: Clinical terminology. This subtrack deals with the translation of clinical terminology. Translating clinical terminology is very relevant due to the existence of many established concepts and multi-word expressions (MWE) that need to be translated not only correctly but also consistently. Systems able to consider not only full sentences but also specific terms are able to provide more accurate translations, something fundamental in the clinical domain.

The gold standard terms were extracted from biomedical literature and electronic health records using information retrieval systems, filtered and translated and revised by professional medical translators. Amongst other semantic classes, the selected terms include diseases, symptoms and findings, procedures, drugs and species. The background set includes over 200,000 concepts in Spanish from the same sources.

The dataset includes a total of 19,128 terms (7,000 for the dev set and 12,128 for the test set). The direction of this sub-track is es2en.

Sub-track 3: Ontology concepts. This sub-track deals with the translation of concepts extracted from ontologies. Ontologies are one of the main ways of structuring knowledge. In the clinical domain, they are widely used mainly to normalize the content of electronic health records. However, their everyday use can be greatly limited by their unavailability in languages other than English. MT systems specifically trained for this type of data can be of great help to improve the impact of these ontologies or to ease a manual translation process.

The gold standard for this task is made up of concepts extracted from various free-access biomedical ontologies and taxonomies and then manually translated by a professional medical translator. Due to their origin, these concepts may present different challenges than terms extracted from free text, such as semi-structured concepts. The background set includes 300,000 concepts in English extracted from the same sources.

The dataset includes a total of 2,189 concepts (300 for the dev set and 1,789 for the test set). The direction of this sub-track is en2es.

3 Baselines

The baselines for en2de, en2fr, en2es, en2pt, de2en, fr2en, es2en, and pt2en were computed using models we trained ourselves in the previous years using Marian NMT (Junczys-Dowmunt et al., 2018). The baselines for en2zh, en2it, en2ru, zh2en, it2en, zh2en were computed using pre-trained Marian models distributed as HuggingFace "Transformers" library models,[8] without trying to increase their performance on the biomedical texts through further fine-tuning. The computation was performed on a single Nvidia A5000 GPU card.

The baselines are strongly outperformed by the participants of the biomedical task, with the exception of **en2it** where all reach similar and very high levels, in excess of 47 BLEU. Especially our zh2en baseline needs improvement.

[8] https://huggingface.co/Helsinki-NLP

4 Teams and systems

In this section we describe the teams and the number of submissions that we received from our two submission systems. When considering both the MEDLINE and the ClinSpEn sub-task, we had a total of 40 participating teams. We describe the submissions for each of them below.

4.1 MEDLINE participation

This year, we received a total of 609 submissions from 37 teams (see Table 3), from the following countries: China (7), France (2), Poland (1), Russia (1), and South Korea (1). Most teams (N=25), however, did not report a country of affiliation.

The number of submissions for each of the MEDLINE test sets are split into to parts: from English in Table 4 and into English in Table 5. We received around 100 more submissions for the test sets into English (354 vs. 255).

As in the 2020 and 2021 editions, we asked participants to fill out a survey with key information regarding the specific material and methods used in their self-identified primary runs used for manual evaluation. The survey comprised 15 questions covering the translation methods and corpora used. For consistency with previous years, the only change to the questionnaire was the addition of a question regarding the method used by teams to estimate the environmental impact of their experiments. We included the CO2 measurement methods identified in (Bannour et al., 2021) as options.

Only six teams supplied information about their "best run", and none reported measuring the environmental impact of their participation to the task. On average, the time spent by participants to supply information for one language pair was 7 minutes and 13 seconds (median: 3 minutes and 27 seconds). This is consistent with the previous survey statistics and suggests that the time commitment for supplying this information is limited, even for teams addressing more than one language pair.

All teams used transformer-based neural MT (NMT), relying mostly on existing implementations. Contrarily to last year, teams addressing several language pairs adapted their setup across them. See Table 6 for details of the teams' methods.

For in-domain data, teams used the training data distributed as part of the task as well as many of the sources described in (Névéol et al., 2018). Additional corpora used for Chinese were prepared by the teams but are not always available or described in detail, except for ParaMed, which relied on the New England Journal of Medicine to create a parallel corpus (Liu and Huang, 2021). Terminologies used by team Summer are available online.[9] The in-domain monolingual corpora used often use different selections of MEDLINE. We can also notice that the use or pre-processing of the same resources can differ between teams as the size reported for seemingly similar data can differ significantly. Table 7 provides details of the in-domain data used by the teams.

For relevant language pairs, parallel data from other WMT tracks (e.g. General or News Task) was used. Out-of-domain data was also used in the form of pre-trained base models. Table 8 shows details of the out-of-domain data used by the teams.

4.2 ClinSpEn Participation

In total, 11 different teams both from academia and industry registered for the ClinSpEn subtask, although only 5 teams ended up submitting their predictions. Four of them participated in all subtracks, with one of them participating only in subtrack 2 (clinical terminology translation). Table 9 presents an overview of the teams who submitted their predictions to the task.

5 Automatic evaluation

In this section we present the automatic evaluation that we performed for the MEDLINE and the ClinSpEn test sets.

5.1 MEDLINE test sets

For the MEDLINE test sets, we calculated the BLEU scores in the same way as previous years (Yeganova et al., 2021). We split the runs that we received into three groups: (i) runs to our BioWMT submission system; (ii) runs to the OCELoT Biomedical Task; and (iii) runs to the OCELoT General Task. As already discussed above, the only difference between the test sets in OCELoT and the ones in our submission system is that the sentences are aligned in OCELoT.

Results for runs to our BioWMT submission systems are presented in Tables 10 and 11. Runs for the Biomedical Task in OCELoT are shown in Tables 12 and 13. The run identifiers were mapped to names (e.g. run1, run2), and the mapping is presented in the Appendix (Tables 23 and 24). Finally,

[9] https://github.com/neulab/covid19-datashare/tree/master/parallel/terminologies

Team ID	Institution	Lime Survey	Publication
AISP-SJTU	AI Speech Co. and Shanghai Jiao Tong University, China		
ALMAnaCH-Inria	Inria, France		
aoligei	-		
bhcs-mt	-		
ChicHealth	ChicHealth, China		
DLUT	-		
DTrans	-		
DTranx	-		
ECNU-MT	East China Normal University, China	✓	(Zheng et al., 2022)
eTranslation	European Commission		
GTCOM	-		
Huawei-BabelTar	Huawei Technologies	✓	(Wang et al., 2022)
Huawei-TSC	Huawei Technologies	✓	(Wu et al., 2022)
JDExploreAcademy.Vega-MT	-		
KwaiMT	-		
Lan-BridgeMT	Lan-Bridge, China		
LanguageX	-		
LT22	-		
Manifold	-		
MeteorMan	-		
neunlplab	-		
njupt-mtt	-		
ONLINE-A	-		
ONLINE-B	-		
Online-G	-		
ONLINE-W	-		
ONLINE-Y	-		
OpenNMT	-		
PAHT	-		
PROMT	PROject MT, Russia		
SPECTRANS	Université Paris Cité, France	✓	(Ballier et al., 2022)
SRPOL	Samsung Research, Poland		
SRT	Samsung Research, South Korea	✓	(Choi et al., 2022)
Summer	Tencent, China	✓	(Li et al., 2022)
super_star	-		
szdx	-		
taicangshaxigaozhong	-		
ustc-mt	-		
V2ray	-		

Table 3: List of the participating teams.

due to the large number of teams and runs, we split the General Task runs into various results tables. The from-English submissions are split into two parts in Tables 14 and 15, while the identifier mapping is provided in Tables 25 and 26. Similarly, the into-English submissions are split into two parts in Table 16 and 17, while the identifier mapping is provided in Tables 27 and 28.

In general, the scores were much higher for runs to the BioWMT submission system than for the ones from the OCELoT test sets. All runs for the BioWMT submissions system outperformed our baseline. We did not provide a baseline for the OCELoT test sets.

5.2 ClinSpEn - CodaLab

The ClinSpEn subtask was evaluated in the CodaLab platform (Pavao et al., 2022). CodaLab is an open-source platform for running competitions, with some of its main advantages being automatic scoring and leaderboard building.

ClinSpEn submissions were evaluated using five common MT metrics: COMET (Rei et al., 2020), METEOR (Banerjee and Lavie, 2005), SacreBLEU (Post, 2018), BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). The main metric used for comparison is SacreBLEU, which is the same as OCELoT uses, and the other metrics are given so that participants are able to evaluate their systems from different perspectives. Part of the evaluation scripts were shared by the MedMTEval organizers

Teams	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh	Total
AISP-SJTU	-	-	-	-	-	-	G2	2
ALMAnaCH-Inria	-	-	-	-	-	G2	-	2
aoligei	-	-	-	-	-	-	O2G2	4
bhcs-mt	-	-	-	-	-	-	G4	4
ChicHealth	-	-	-	-	-	-	B1	1
DLUT	-	-	-	-	-	-	G4	4
Dtranx	O1G3	O2	O3	O3	O3	O2G2	O2G2	23
eTranslation	-	-	-	-	-	G3	-	3
ECNU-MT	-	-	-	-	-	-	B1	1
GTCOM	-	-	-	-	-	-	G3	3
Huawei-BabelTar	B3	B3	B3	B3	B3	B3	B3	21
Huawei-TSC	B3O6	-	B3O4	-	-	B3O3	B3O6G7	38
JDExploreAcademy.Vega-MT	G2	-	-	-	-	G2	G7	11
KwaiMT	-	-	-	-	-	-	G3	3
Lan-BridgeMT	O2G4	-	-	-	-	O2G4	O4G4	20
LanguageX	-	-	-	-	-	-	G4	4
Manifold	-	-	-	-	-	-	G7	7
MeteorMan	-	-	-	-	-	-	G1	1
neunlplab	-	-	-	-	-	-	G6	6
njupt-mtt	O1	-	O3	-	-	O3G4	O3G7	21
ONLINE-A	G1	-	-	-	-	G2	G1	4
ONLINE-B	G1	-	-	-	-	G1	G2	4
Online-G	G1	-	-	-	-	G1	G1	3
ONLINE-W	G2	-	-	-	-	G1	G1	4
ONLINE-Y	G2	-	-	-	-	G2	G2	6
OpenNMT	G5	-	-	-	-	-	-	5
PAHT	-	-	-	-	-	-	B1	1
PROMT	G3	-	-	-	-	G5	-	8
SPECTRANS	-	-	O4	-	-	-	-	4
SRPOL	-	-	-	-	-	G6	-	6
SRT	-	B3	-	-	-	-	-	3
super_star	-	-	-	-	-	G2	-	2
szdx	-	-	-	-	-	-	G7	7
taicangshaxigaozhong	-	-	-	-	-	-	G2	2
ustc-mt	-	-	O6	-	-	O2	O2G6	16
V2ray	-	-	-	-	-	-	G1	1
Total	40	8	26	6	6	55	114	255

Table 4: Overview of the submissions from all teams and test sets translating from English. We identify submissions using the WMT Biomedical Submission System (WMTBio) with a "B", the ones for OCELoT Biomedical Task with an "O", and the ones for OCELoT General Task with an "G". The value next to the letter indicates the number of runs for the corresponding test set, language pair, and team.

(Ezhergina et al., 2022), who used the HuggingFace datasets library (Lhoest et al., 2021). Multiple tests were performed to check that the results of our evaluation scripts are comparable to those returned by OCELoT and the WMT submission system. In total, participants were allowed to upload up to 7 predictions for each sub-track.

Tables 18, 19 and 20 show the overall results of each of the three sub-tracks. Only each team's best run is presented.

6 Manual evaluation

For the MEDLINE test sets, we performed a manual evaluation for some selected runs from some of the teams. In this section we describe how the teams and runs were selected, the results of the manual evaluation, and our observations on the quality of the translations.

6.1 Selected teams and submissions

A team qualified for manual evaluation if the participants either submitted a survey or a publications with details about their submission (see Section 4). Only the following six teams complied with this requirement: ECNU-MT, Huawei-BabelTar, Huawei-TSC, SPECTRANS, SRT, and Summer.

During the submission, we asked the participants to identify a primary submission for each language pair, as indicated in Tables 10, 11, 12, 13, 14, 15, 16, and 17. For those teams who submitted runs to both submission systems, we chose the ones sent to the BioWMT submission system. The Huawei-

Teams	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en	Total
AISP-SJTU	-	-	-	-	-	-	G1	1
ALMAnaCH-Inria	-	-	-	-	-	G2	-	2
aoligei	-	-	-	-	-	-	O4G5	9
bhcs-mt	-	-	-	-	-	-	G5	5
bymt	-	-	-	-	-	-	G1	1
ChicHealth	-	-	-	-	-	-	B3	3
Dtranx	O3G3	O1	O3	O3	O3	O2G2	O2G2	24
DLUT	-	-	-	-	-	-	G3	3
ECNU-MT	-	-	-	-	-	-	B2	2
Huawei-BabelTar	B3	B3	B3	B3	B3	B3	B3	21
Huawei-TSC	B3O4	-	B3O3	-	-	B3O4	B3O6G4	33
JDExploreAcademy.Vega-MT	G2	-	-	-	-	G3	G7	12
KwaiMT	-	-	-	-	-	-	G3	3
Lan-BridgeMT	O2G3	-	-	-	-	O2G3	O6G4	20
LanguageX	-	-	-	-	-	-	G6	6
Liaoning University	-	-	-	-	-	-	G3	3
LT22	G5	-	-	-	-	-	-	5
neunlplab	-	-	-	-	-	-	G6	6
njupt-mtt	O1	-	O3	-	-	O2G3	O2G7	18
ONLINE-A	G1	-	-	-	-	G1	G1	3
ONLINE-B	G1	-	-	-	-	G1	G1	3
Online-G	G1	-	-	-	-	G1	G1	3
ONLINE-W	G2	-	-	-	-	G1	G1	4
ONLINE-Y	G2	-	-	-	-	G2	G2	6
PAHT	-	-	-	-	-	-	B1	1
pingan_mt	-	-	-	-	-	-	G1	1
PROMT	G2	-	-	-	-	G1	-	3
SRPOL	-	-	-	-	-	G7	-	7
SPECTRANS	-	-	O4	-	-	-	-	4
SRT	-	B3	-	-	-	-	-	3
star	-	-	-	-	-	-	G4	4
super_star	-	-	-	-	-	-	G6	6
szdx	-	-	-	-	-	-	O1G7	8
Summer	-	-	-	-	-	-	B3	3
taicangshaxigaozhong	-	-	-	-	-	-	G4	4
ustc-mt	-	-	O5	-	-	O2	O1G5	13
V2ray	-	-	-	-	-	-	G1	1
Total	38	7	62	6	6	107	128	354

Table 5: Overview of the submissions from all teams and test sets translating into English. We identify submissions using the WMT Biomedical Submission System (WMTBio) with a "B", the ones for OCELoT Biomedical Task with an "O", and the ones for OCELoT General Task with an "G". The value next to the letter indicates the number of runs for the corresponding test set, language pair, and team.

Team ID	Language pair	NMT implementation	Trained	Fine-Tuned	BT	LM
ECNU_MT	en2zh, zh2en	fairseq	No	Yes	Yes	Yes
Huawei_BabelTar	en/de,es,fr,zh	fairseq	No	Yes	Yes, into en	No
Huawei_BabelTar	en/it	fairseq	No	Yes	Yes, for it2en	Yes, for it2en
Huawei_BabelTar	en/pt,ru	fairseq	No	Yes	Yes, from en	Yes, into en
Huawei_TSC	en/ru	Fairseq	No	Yes	Yes	No
Huawei_TSC	en/de, en/zh	Marian, Fairseq	No	Yes	Yes	No
Huawei_TSC	en/fr	Marian, Fairseq	Yes	No	Yes	No
SPECTRANS	en2fr	SYSTRAN Pure Neural Server 9.8	No	Yes	No	Yes
SRT	es2en	Fairseq	Yes	No	Yes	No
Summer	zh2en	Fairseq	Yes	No	Yes	No

Table 6: Overview of methods used by participating teams. Information is self-reported through the dedicated survey for each selected "best run". BT indicates if backtranslation is used and LM if language models were used.

Language pair	team	Parallel corpus	size (sentence pairs)	Monolingual corpus	size (sentences)
de/en	Huawei_BabelTar	MEDLINE corpus supplied by WMT biomedical task organizers	2.4 M	Yes	53 M (en)
	Huawei_TSC	UFAL corpus and "internal corpus"	2.75M	No	-
es/en	Huawei_BabelTar	MEDLINE corpus supplied by WMT biomedical task organizers	1.1 M	Yes	52.5 M (en)
	Huawei_TSC	corpus provided by WMT biomedical task organizers	8.1 M	Yes	8 M (en)
	SRT	MEDLINE, UFAL, MeSpEN and Scielo	3.47 M	Yes	3.5M (es), 13.9M (en)
fr/en	Huawei_BabelTar	MEDLINE corpus supplied by WMT biomedical task organizers	2.8 M	Yes	53 M (en)
	Huawei_TSC	corpus provided by WMT biomedical task organizers	6 M	Yes	2 M (en) 45M (en)
	SPECTRANS	in-house translation memory on diabetes and UFAL	2,700 (TM)	No	-
it/en	Huawei_BabelTar	MEDLINE corpus supplied by WMT biomedical task organizers	139 K	Yes	55 M (en)
pt/en	Huawei_BabelTar	MEDLINE corpus supplied by WMT biomedical task organizers	7.1 M	Yes	52.5 M (en)
en/ru	Huawei_BabelTar	Corpus supplied by WMT biomedical task organizers.	32 K	Yes	52.5 M (en)
	Huawei_TSC	Corpus supplied by organizers	24 K	Yes	46 M (en)
en/zh	ECNU_MT	NEJM en-zh corpus	66 K	Yes	40 M (en)
	Huawei_BabelTar	TAUS corpus	847 K	Yes	53 M (en)
	Huawei_TSC	UFAL and in-house corpus (unspecified)	10.87M	No	-
	Summer	MEDLINE, TAUS and covid-19 terminology by Google and Facebook	0.5 M	Yes	6.9 M (en)

Table 7: Overview of in-domain corpora used by participating teams. Information is self reported through our survey for each selected "best run" (information on the NVIDIA model is inferred from their task paper).

Language pair	team	Parallel corpus	size (sentence pairs)	Monolingual corpus	size (sentences)
en/de	Huawei_BabelTar	"in house data"	6 M	No	-
	Huawei_TSC	WMT general corpus and "internal corpus"	200 M	Yes	10M (de) 46M (en)
en/es	Huawei_BabelTar	WikiMatrix	3.3 M	No	-
	Huawei_TSC	WMT general corpus and "internal corpus"	200 M	No	-
	SRT	ParaCrawl, CommonCrawl, Europarl, News Commentary, Tatoeba, and UN Corpus	518 M	No	-
en/fr	Huawei_BabelTar	"in house corpus"	3 M	No	-
	Huawei_TSC	"in house data"	600 M	No	-
	SPECTRANS	UFAL Corpus	2,7 M	No	-
en/it	Huawei_BabelTar	"in house data"	6 M	No	-
en/pt	Huawei_BabelTar	WikiMatrix	3 M	No	-
en/ru	Huawei_BabelTar	"in house data"	3 M	No	-
	Huawei_TSC	Corpus supplied by the WMT 2022 general task	200 M	Yes	46 M (en) 40 M (ru)
en/zh	ECNU_MT	NA	-	No	-
	Huawei_BabelTar	"in house corpus"	3 M	No	-
	Huawei_TSC	"in house data"	200 M	Yes	46M (en) 92M (zh)
	Summer	Corpus supplied by the WMT 2021 News task	30.6 M	Yes	132 M (en)

Table 8: Overview of out-of-domain (OOD) corpora used by participating teams. Information is self reported through our survey for each selected "best run".

Team ID	Affiliation	Clinical Cases (en2es)	Terminology (es2en)	Ontology (en2es)
Avellana Translation	Avellana Translation	✓	✓	✓
DtranX	DtranX	✓	✓	✓
Huawei	Huawei Technologies		✓	
Logrum_UoM	University of Manchester	✓	✓	✓
Optum	Optum	✓	✓	✓

Table 9: List of the participating teams who submitted results to the ClinSpEn subtask.

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh
ChicHealth	run1	-	-	-	-	-	-	55.71
ECNU-MT	run1	-	-	-	-	-	-	39.85
HuaweiTSC	run1	39.00	-	40.17*	-	-	41.27	50.79
	run2	39.14*	-	38.81	-	-	40.53	50.78*
	run3	38.91	-	39.00	-	-	40.63*	50.68
Huawei-BabelTar	run1	33.42	44.70	37.85	46.49	52.55	36.97	47.68
	run2	33.13	44.15	37.49	47.83	51.74	36.74	47.30
	run3	33.04	44.75	36.21	48.48	51.47	37.03	45.13
PAHT	run1	-	-	-	-	-	-	48.26
SRT	run1	-	52.14	-	-	-	-	-
	run2	-	51.96*	-	-	-	-	-
	run3	-	52.35	-	-	-	-	-
Baseline	-	29.43	39.15	28.12	47.13	42.39	27.59	39.79

Table 10: BLEU scores for "OK" aligned MEDLINE test sentences, from English, for submissions to the BioWMT Biomedical system. Primary runs are marked by *.

Teams	Runs	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en
ChicHealth	run1	-	-	-	-	-	-	34.27
	run2	-	-	-	-	-	-	36.48
	run3	-	-	-	-	-	-	46.14*
ECNU-MT	run1	-	-	-	-	-	-	24.75*
	run2	-	-	-	-	-	-	24.49
HuaweiTSC	run1	46.95	-	50.95*	-	-	48.86	42.69
	run2	47.12*	-	50.36	-	-	50.01*	42.56*
	run3	46.82	-	50.48	-	-	49.58	42.76
Huawei-BabelTar	run1	43.10	56.60	49.08	48.83	56.03	46.16	46.12
	run2	43.75	59.02	48.86	49.16	55.44	46.26	42.49
	run3	43.38	58.64	49.36	49.89	55.63	46.75	41.80
PAHT	run1	-	-	-	-	-	-	31.16
SRT	run1	-	59.54	-	-	-	-	-
	run2	-	59.43	-	-	-	-	-
	run3	-	60.45*	-	-	-	-	-
Summer	run1	-	-	-	-	-	-	44.39*
	run2	-	-	-	-	-	-	44.31
	run3	-	-	-	-	-	-	46.17
Baseline	-	33.28	40.42	37.29	42.98	47.57	31.23	20.41

Table 11: BLEU scores for “OK” aligned MEDLINE test sentences, into English, for submissions to the BioWMT Biomedical system. Primary runs are marked by *.

BabelTar did not not indicate their primary run for some languages, and so we chose the ones with the highest scores, namely: run1 for en2de, en2fr, pt2en, and en2pt; run2 for de2en and es2en; and run3 for en2es, fr2en, it2en, en2it, ru2en, en2ru, zh2en, en2zh.

For submissions into English, we randomly selected the abstracts until we achieved at least 100 perfectly aligned (OK) sentences (see Table 2). We performed pairwise comparison between the reference translation and the selected submissions. The results from the manual validation are presented in Table 21. Unfortunately, we could not perform manual validation for submissions for de2en and

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh
aoligei	run1	-	-	-	-	-	-	38.71
	run1	-	-	-	-	-	-	38.25*
Dtranx	run1	34.84*	49.18*	34.73	48.92	47.83	30.78*	41.14
	run2	-	49.18	35.18	47.52	37.84	17.45	36.25*
	run3	-	-	23.84*	29.20*	24.44*	-	-
Huawei-TSC	run1	34.15	-	35.02	-	-	26.88	40.25
	run2	34.04	-	34.98	-	-	30.59	40.13
	run3	34.28	-	35.56	-	-	26.73*	40.21
	run4	33.97	-	36.13*	-	-	-	39.99
	run5	34.28	-	-	-	-	-	40.12
	run6	34.28*	-	-	-	-	-	39.42
Lan-BridgeMT	run1	31.10	-	-	-	-	25.52*	37.86
	run2	31.67*	-	-	-	-	25.28	36.90
	run3	-	-	-	-	-	-	37.99
	run4	-	-	-	-	-	-	37.98*
njupt-mtt	run1	33.94	-	35.41	-	-	25.64	36.53
	run2	-	-	35.07	-	-	27.09	40.25
	run3	-	-	34.69	-	-	26.73	39.87
SPECTRANS	run1	-	-	20.68	-	-	-	-
	run2	-	-	31.63*	-	-	-	-
	run3	-	-	7.32	-	-	-	-
	run4	-	-	20.34	-	-	-	-
ustc-mt	run1	-	-	33.69	-	-	26.97	40.02
	run2	-	-	34.40	-	-	30.95	39.63
	run3	-	-	35.30	-	-	-	-
	run4	-	-	34.91	-	-	-	-
	run5	-	-	35.41	-	-	-	-
	run6	-	-	35.55	-	-	-	-

Table 12: BLEU scores for the OCELoT Biomedical Task, from English. An asterisk * indicates the primary run.

it2en.

For submissions from English, we manually selected 19 sentences from one of the clinical case reports, namely, PMID 35144678. Subsequently, we completed the sets with abstracts from the respective test sets. For the abstracts and exclusively for en2fr, for which a reference translation for the clinical case reports is available, we carry out a pairwise comparison between the reference translation and the selected submissions. For the case report for the remaining languages, we could only perform pairwise comparisons between teams' submissions. The results from the manual validation are presented in Table 22.

In both tables, we show in bold the comparisons in which one of the teams (or the reference translation) was statistically significant, according to the Wilcoxon test. The reference translation had a similar quality to many of the submissions. However, none of the teams was (statistically significant) superior than the reference translation.

6.2 Quality of the translations

Here we discuss the quality of the translations after manual validation of the selected abstracts and clinical case report.

en2fr As in previous years, the overall translation quality was high, with many automatically produced sentences exhibiting only small differences with the reference translation. In the examples shown below, correct translations are shown in black font while incorrect ones appear in red fond. Passages underlined within the same example block mark text that should carry the same meaning across statements.

(1) **en:** risk of short-term stroke
fr$_1$: risque d'AVC à court terme
fr$_2$: risque d'accident vasculaire cérébral de courte durée

(2) **en:** the long-term stroke ARD
fr$_1$: la DRA de l'AVC à long terme
fr$_2$: la maladie d'Alzheimer et les démences apparentées à long terme

However, longer or more complex sentences seemed more difficult to address for automatic systems. For examples, acronym modifiers were sometimes translated erroneously 1. We also noticed recurring issues pertaining to acronym translation (Example 2) as well as consistency throughout an entire document.

Teams	Runs	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en
aoligei	run1	-	-	-	-	-	-	40.85
	run2	-	-	-	-	-	-	39.75
	run3	-	-	-	-	-	-	41.26*
	run4	-	-	-	-	-	-	40.24
DTranx	run1	35.55	54.21*	44.94	45.85	54.89	38.40	41.27
	run2	22.60*	-	46.38	42.04	53.67	19.34*	39.22*
	run3	37.28	-	26.91*	25.28*	28.06*	-	-
Huawei-TSC	run1	37.59	-	45.66	-	-	35.57	41.25
	run2	37.49	-	51.86	-	-	36.33	41.50*
	run3	37.62	-	46.76	-	-	35.85	41.33
	run4	37.60*	-	-	-	-	36.33*	41.33
	run5	-	-	-	-	-	-	41.66
	run6	-	-	-	-	-	-	41.46
Lan-BridgeMT	run1	35.09	-	-	-	-	31.71*	40.64
	run2	34.99*	-	-	-	-	31.24	40.09
	run3	-	-	-	-	-	-	39.78
	run4	-	-	-	-	-	-	39.31
	run5	-	-	-	-	-	-	40.73*
njupt-mtt	run1	37.09	-	45.68	-	-	35.05	41.39
	run2	-	-	44.85	-	-	35.87	41.32
	run3	-	-	44.94	-	-	-	-
SPECTRANS	run1	-	-	25.81	-	-	-	-
	run2	-	-	40.10*	-	-	-	-
	run3	-	-	25.87	-	-	-	-
	run4	-	-	9.69	-	-	-	-
ustc-mt	run1	-	-	45.11	-	-	35.39	41.05
	run2	-	-	44.81	-	-	38.48	-
	run3	-	-	45.77	-	-	-	-
	run4	-	-	45.27	-	-	-	-
	run5	-	-	0.02	-	-	-	-
szdx	-	-	-	-	-	-	-	36.00

Table 13: BLEU scores for the OCELoT Biomedical Task, into English. An asterisk * indicates the primary run.

For example, the acronym *POAF*, corresponding to the term *Perioperative atrial fibrillation*, was translated as *POAF*, *FOPA*, *FPO* or *FAPO*. Systems commonly used a combination of two or more of these solutions throughout a whole document, while the reference translation consistently used the correct translation, *FAPO*.

This year, manual validation for en2fr was performed by one evaluator with translation training and one clinician. The overall agreement on individual pair comparison was moderate at 64%. However, the overall ordering of systems and reference according to both annotator remained unchanged.

fr2en As in previous years, translation quality was high, resulting in many automatically produced translations whose quality was indistinguishable from that of reference translations. Concerning the quality of this year's references, they generally corresponded better to direct (as opposed to approximate) translations of the source abstracts, with respect to previous years. This is reflected by the pairwise comparison, which shows that the reference translation is systematically preferred over automatic translations. The most common translation errors were in term and acronym translation (Examples 3-6), prepositional and adjectival attachment (Examples 7 and 8 and in lack of capitalisation (of terms and in particular of acronyms). Term translation was particularly important for overall translation quality, often counterbalancing other more minor errors such as the naturalness of lexical and syntactic choices and correct capitalisation.

(3) **fr:** polyradiculonévrite inflammatoire démyélinisante chronique
en$_1$: chronic inflammatory demyelinating polyradiculoneuropathy
en$_2$: *chronic inflammatory demyelinating polyradiculoneuritis

(4) **fr:** défaut de croissance staturo-pondérale
en$_1$: failure to thrive
en$_2$: *staturo-weight growth defect

(5) **fr:** les inhibiteurs des cotransporteurs sodium-glucose de type 2 (iSGLT2, gliflozines)
en$_1$: sodium-glucose cotransporter type 2 inhibitors (SGLT2i, gliflozins)

Teams	Runs	en2de	en2ru	en2zh
AISP-SJTU	run1	-	-	37.74
	run2	-	-	37.70
ALMAnaCH-Inria	run1	-	20.22	-
	run2	-	9.77	-
aoligei	run1	-	-	38.71
	run2	-	-	38.25
bhcs-mt	run1	-	-	33.61
	run2	-	-	34.34
	run3	-	-	39.82
	run4	-	-	39.82
DLUT	run1	-	-	36.22
	run2	-	-	35.58
	run3	-	-	36.32
	run4	-	-	30.54
Dtranx	run1	0.03	30.78	41.14
	run2	34.84	17.45	37.98
	run3	34.43	-	- -
eTranslation	run1	-	27.53	-
	run2	-	27.28	-
	run3	-	27.53	-
GTCOM	run1	-	-	38.18
	run2	-	-	37.06
	run3	-	-	36.94
HuaweiTSC	run1	-	-	36.36
	run2	-	-	35.72
	run3	-	-	35.89
	run4	-	-	37.95
	run5	-	-	35.66
	run6	-	-	37.95
	run7	-	-	39.42
JDExploreAcademy.Vega-MT	run1	33.32	29.77	39.24
	run2	33.50	29.49	41.16
	run3	-	-	41.16
	run4	-	-	41.16
	run5	-	-	40.40
	run6	-	-	40.63
	run7	-	-	39.82
KwaiMT	run1	-	-	37.34
	run2	-	-	41.06
	run3	-	-	41.06
Lan-Bridge	run1	31.10	25.52	37.86
	run2	31.67	25.28	37.86
	run3	31.84	25.38	36.90
	run4	34.43	30.91	37.97
LanguageX	run1	-	-	42.17
	run2	-	-	41.79
	run3	-	-	41.35
	run4	-	-	41.57
Manifold	run1	-	-	38.00
	run2	-	-	38.40
	run3	-	-	37.99
	run4	-	-	38.10
	run5	-	-	38.15
	run6	-	-	38.21
	run7	-	-	38.31
MeteorMan	run1	-	-	38.58
neunlplab	run1	-	-	34.76
	run2	-	-	35.21
	run3	-	-	35.21
	run4	-	-	35.03
	run5	-	-	35.14
	run6	-	-	35.30

Table 14: BLEU scores for OCELoT General Task, from English (part 1/2).

Teams	Runs	en2de	en2ru	en2zh
njupt-mtt	run1	-	26.43	40.25
	run2	-	27.20	36.53
	run3	-	30.95	41.16
	run4	-	25.36	37.19
	run5	-	-	37.09
	run6	-	-	37.03
	run7	-	-	37.99
ONLINE-A	run1	33.21	28.04	37.94
	run2	-	28.04	-
ONLINE-B	run1	34.88	30.90	41.17
	run2	-	-	41.17
Online-G	run1	33.76	29.68	37.31
ONLINE-W	run1	34.88	31.59	39.42
	run2	37.37	-	-
ONLINE-Y	run1	34.88	30.90	41.17
	run2	33.38	28.23	37.79
OpenNMT	run1	30.72	-	-
	run2	30.92	-	-
	run3	30.47	-	-
	run4	29.48	-	-
	run5	30.89	-	-
PROMT	run1	32.82	29.18	-
	run2	32.70	31.13	-
	run3	32.70	31.07	-
	run4	-	29.68	-
	run5	-	29.18	-
SRPOL	run1	-	27.78	-
	run2	-	27.61	-
	run3	-	27.24	-
	run4	-	27.62	-
	run5	-	27.52	-
	run6	-	27.58	-
super_star	run1	-	-	36.94
	run2	-	-	41.06
szdx	run1	-	-	38.58
	run2	-	-	38.23
	run3	-	-	38.25
	run4	-	-	38.25
	run5	-	-	38.25
	run6	-	-	38.25
	run7	-	-	38.25
taicangshaxigaozhong	run1	-	-	13.75
	run2	-	-	38.58
ustc-mt	run1	-	-	36.45
	run2	-	-	32.60
	run3	-	-	31.31
	run4	-	-	38.01
	run5	-	-	35.46
	run6	-	-	38.45
V2ray	run1	-	-	41.16

Table 15: BLEU scores for OCELoT General Task, from English (part 2/2).

en$_2$: *type 2 sodium glucose co-transporter inhibitors (iSGLT2, gliflozins)

(6) **fr:** une VCE pour OGIB en pratique courante
en$_1$: VCE for OGIB in routine practice
en$_2$: *an ECV for OGIB in current practice[10]

(7) **fr:** pour les migraines et céphalées en grappe
en$_1$: for migraines and cluster headaches
en$_2$: *for cluster migraines and headaches

(8) **fr:** les personnes non diabétiques
en$_1$: non-diabetic people
en$_2$: *non-people with diabetes

As an additional comment, some of the MT out-

[10]This example is interesting, since the original French uses English acronyms rather than French ones, presumably as they are well-known terms that have been borrowed into scientific French. The correct English translation is therefore to use the same acronyms as the French.

Teams	Runs	de2en	ru2en	zh2en
AISP-SJTU	run1	-	-	39.22
ALMAnaCH-Inria	run1	-	25.64	-
	run2	-	21.69	-
aoligei	run1	-	-	40.85
	run2	-	-	41.45
	run3	-	-	40.03
	run4	-	-	40.34
	run5	-	-	40.85
bhcs-mt	run1	-	-	31.75
	run2	-	-	39.09
	run3	-	-	39.46
	run4	-	-	40.95
	run5	-	-	41.03
bymt	run1	-	-	39.22
Dtranx	run1	35.55	38.40	41.27
	run2	37.28	19.34	39.22
	run3	22.60	-	-
DLUT	run1	-	-	33.10
	run2	-	-	32.95
	run3	-	-	33.22
HuaweiTSC	run1	-	-	36.85
	run2	-	-	34.63
	run3	-	-	36.73
	run4	-	-	36.73
JDExploreAcademy.Vega-MT	run1	35.92	37.90	39.03
	run2	36.24	37.85	40.63
	run3	-	37.90	40.73
	run4	-	-	40.48
	run5	-	-	41.14
	run6	-	-	41.41
	run7	-	-	41.27
KwaiMT	run1	-	-	41.09
	run2	-	-	39.89
	run3	-	-	39.88
Lan-Bridge	run1	35.09	31.71	40.64
	run2	34.99	31.24	40.37
	run3	35.62	38.86	40.31
	run4	-	-	40.73
LanguageX	run1	-	-	41.95
	run2	-	-	39.50
	run3	-	-	41.21
	run4	-	-	40.57
	run5	-	-	41.38
	run6	-	-	41.08
Liaoning University	run1	-	-	39.44
	run2	-	-	34.62
	run3	-	-	34.67
LT22	run1	24.69	-	-
	run2	24.59	-	-
	run3	24.22	-	-
	run4	23.19	-	-
	run5	23.19	-	-
neunlplab	run1	-	-	34.76
	run2	-	-	35.21
	run3	-	-	35.21
	run4	-	-	35.03
	run5	-	-	35.14
	run6	-	-	35.30

Table 16: BLEU scores for OCELoT General Task, into English (part 1/2).

puts appeared robust to unexpected variation in the source texts, such as rare cases of odd capitalisation, additional spaces within words and the use of inclusive writing, as can be seen in Example 5 with the word *patient.e.s* 'patient (m/f)', indicating the masculine and feminine forms simultaneously.

Teams	Runs	de2en	ru2en	zh2en
njupt-mtt	run1	-	35.53	34.51
	run2	-	35.66	41.38
	run3	-	33.09	34.56
	run4	-	-	35.63
	run5	-	-	36.40
	run6	-	-	0.5
	run7	-	-	35.05
ONLINE-A	run1	35.76	36.72	36.67
ONLINE-B	run1	35.50	38.27	41.03
Online-G	run1	35.30	37.69	35.88
ONLINE-W	run1	35.50	32.51	37.41
	run2	37.62	-	-
ONLINE-Y	run1	35.50	38.27	41.03
	run2	35.64	36.05	36.89
pingan_mt	run1	-	-	41.86
PROMT	run1	35.06	33.10	-
	run2	35.06	-	-
SRPOL	run1	-	33.68	-
	run2	-	34.22	-
	run3	-	33.77	-
	run4	-	34.54	-
	run5	-	34.58	-
	run6	-	34.83	-
	run7	-	34.85	-
star	run1	-	-	41.26
	run2	-	-	41.71
	run3	-	-	40.20
	run4	-	-	40.85
super_star	run1	-	-	40.42
	run2	-	-	39.48
	run3	-	-	41.07
	run4	-	-	41.59
	run5	-	-	38.80
	run6	-	-	40.85
szdx	run1	-	-	36.00
	run2	-	-	39.22
	run3	-	-	39.20
	run4	-	-	39.22
	run5	-	-	39.22
	run6	-	-	11.95
	run7	-	-	39.22
taicangshaxigaozhong	run1	-	-	39.22
	run2	-	-	39.22
	run3	-	-	14.77
	run4	-	-	39.22
ustc-mt	run1	-	-	23.17
	run2	-	-	25.00
	run3	-	-	34.96
	run4	-	-	35.71
	run5	-	-	36.68
V2ray	run1	-	-	41.17

Table 17: BLEU scores for the OCELoT General Task, into English (part 2/2).

Teams	Run	COMET	METEOR	SacreBLEU	BLEU	ROUGE
Avellana Translation	run1	0.392	0.643	36.64	35.19	0.633
DtranX	run1	0.461	0.663	**41.06**	39.36	0.649
Logrus_UoM	run1	0.423	0.633	38.17	36.50	0.627
Optum	run4	0.442	0.644	38.12	36.42	0.628

Table 18: Results for the first ClinSpEn sub-track (en2es clinical case report translation).

Teams	Run	COMET	METEOR	SacreBLEU	BLEU	ROUGE
Avellana Translation	run1	0.196	0.570	15.88	15.65	0.686
DtranX	run1	1.115	0.611	35.84	35.21	0.701
Huawei	run7	1.190	0.624	**41.57**	41.32	0.721
Logrus_UoM	run1	0.979	0.588	26.87	26.67	0.671
Optum	run2	0.982	0.574	27.94	27.57	0.656

Table 19: Results for the second ClinSpEn sub-track (es2en clinical terminology translation).

Teams	Run	COMET	METEOR	SacreBLEU	BLEU	ROUGE
Avellana Translation	run1	0.384	0.570	31.72	30.42	0.762
DtranX	run1	1.249	0.627	**58.24**	57.24	0.783
Logrus_UoM	run1	0.949	0.626	39.10	36.74	0.768
Optum	run1	1.119	0.588	44.97	43.96	0.747

Table 20: Results for the third ClinSpEn sub-track (en2es ontology concept translation).

Lang. dir.	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
es2en	**reference** vs. Huawei-BabelTar	14	**6**	4	4	106	23	**47**	36
	reference vs. SRT	14	3	3	8	106	14	45	**47**
	Huawei-BabelTar vs. SRT	14	3	7	4	106	11	73	22
fr2en	SPECTRANS vs. Huawei-TSC	18	5	0	13	103	23	34	**46**
	SPECTRANS vs. **Huawei-BabelTar**	18	3	2	**13**	103	24	33	**46**
	SPECTRANS vs. **reference**	18	3	0	**15**	103	23	12	**68**
	Huawei-TSC vs. Huawei-BabelTar	18	11	3	4	103	40	44	**19**
	Huawei-TSC vs. reference	18	6	1	11	103	35	24	44
	Huawei-BabelTar vs. **reference**	18	2	5	**11**	103	29	22	**52**
pt2en	Huawei-BabelTar vs. reference	12	1	10	1	101	18	70	13
ru2en	reference vs. Huawei-BabelTar	14	**7**	6	1	108	31	59	18
	reference vs. Huawei-TSC	14	8	3	3	108	**44**	54	10
	Huawei-BabelTar vs. Huawei-TSC	14	1	11	2	108	7	87	14
zh2en	Summer vs. Huawei-BabelTar	17	**12**	2	3	-	-	-	-
	Summer vs. reference	17	6	7	4	-	-	-	-
	Summer vs. Huawei-TSC	17	2	11	4	-	-	-	-
	Summer vs. ECNU-MT	17	14	2	1	-	-	-	-
	Huawei-BabelTar vs. reference	17	1	9	**7**	-	-	-	-
	Huawei-BabelTar vs. Huawei-TSC	17	1	4	**12**	-	-	-	-
	Huawei-BabelTar vs. ECNU-MT	17	11	0	6	-	-	-	-
	reference vs. Huawei-TSC	17	4	9	4	-	-	-	-
	reference vs. ECNU-MT	17	**16**	1	0	-	-	-	-
	Huawei-TSC vs. ECNU-MT	17	**14**	3	0	-	-	-	-

Table 21: Pairwise manual evaluation results for the MEDLINE abstracts test set (into English). We show in bold the values which were statistically significant (Wilcoxon test). We only show the team (or reference) in bold, if both the abstracts and sentences were statistically significant (bold).

Nevertheless, most systems struggled to deal with the ambiguity linked to the translation of personal pronouns *sa*, *son*, *ses* 'his/her' in a context where it refers to an unspecified individual (e.g. *the teenager, the child*, etc.); most systems chose the masculine 'his', whereas the correct translation would either be gender neutral 'they' or 'his or her'.

From the manual evaluation results (cf. Table 21), it appears that Huawei-TSC is the superior system; although results are not significant for comparisons against the other two systems), it is the only system of the three that is not significantly worse than the reference translation. Results for abstracts and for sentences appear to correlate, although it was possible on occasions for an abstract to be of better quality than another despite having fewer better individual sentences (due to the differing importance of different errors).

Lang. dir.	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
en2de	reference vs. Huawei-TSC	11	2	5	4	79	12	50	17
	reference vs. Huawei-BabelTar	11	**10**	0	1	79	**57**	20	2
	Huawei-TSC vs. Huawei-BabelTar	12	**9**	3	0	96	**70**	24	2
en2es	Huawei-BabelTar vs. **SRT**	11	1	2	**8**	115	11	35	**57**
	reference vs. Huawei-BabelTar	11	**7**	2	1	86	**51**	25	10
	reference vs. SRT	10	0	7	3	86	16	50	20
en2fr	**reference** vs SPECTRANS	6	**6**	0	0	87	**79**	7	0
	reference vs. Huawei-TSC	6	**6**	0	0	87	**76**	10	0
	reference vs. Huawei-BabelTar	6	**6**	0	0	87	**75**	10	1
	SPECTRANS vs. Huawei-TSC	6	1	1	4	87	27	20	40
	SPECTRANS vs. Huawei-BabelTar	6	**5**	1	0	87	**63**	18	6
	Huawei-TSC vs. Huawei-BabelTar	6	**6**	0	0	87	**59**	25	3
en2it	Huawei-BabelTar vs. reference	11	3	3	5	100	18	56	26
en2pt	reference vs. Huawei-BabelTar	6	0	6	5	105	19	54	32
en2ru	Huawei-TSC vs. Huawei-BabelTar	9	3	4	2	102	15	66	16
	Huawei-TSC vs. reference	8	3	2	3	84	14	56	13
	Huawei-BabelTar vs. reference	8	2	2	4	84	15	55	13
en2zh	Huawei-BabelTar vs. ECNU-MT	14	8	3	3	-	-	-	-
	Huawei-BabelTar vs. Huawei-TSC	14	10	1	3	-	-	-	-
	Huawei-BabelTar vs. reference	13	3	2	8	-	-	-	-
	ECNU-MT vs. Huawei-TSC	14	7	4	3	-	-	-	-
	ECNU-MT vs. reference	13	2	5	6	-	-	-	-
	Huawei-TSC vs. reference	13	2	1	**10**	-	-	-	-

Table 22: Pairwise manual evaluation results for the MEDLINE abstracts test set (from English). We show in bold the values which were statistically significant (Wilcoxon test). We only show the team (or reference) in bold, if both the abstracts and sentences were statistically significant (bold).

en2pt As shown in Table 22, the translations from the Huawei-BabelTar team achieved a similar quality as the reference translation. Similar to previous years, the translations had a good quality and we found just some few mistakes. For instance, errors in acronyms are still present, e.g. “Reforma Psiquiátrica Brasileira (RBP)” instead of “Reforma Psiquiátrica Brasileira (RPB)”. Some translations might not include mistakes, but we thought that one of them was clearer than the other, e.g. “desfechos desfavoráveis tanto para a mãe quanto para o feto” (unfavorable outcomes for both mother and fetus) instead of “maus desfechos maternos e fetais” (poor maternal and fetal outcomes). Finally, we found it interesting that all query terms remained in English, namely “status epilepticus”, “refractory”, “treatment” and “topiramate”, for both translations, in one particular sentence that discussed queries to a search tool.

pt2en As shown in Table 21, the translations from the Huawei-BabelTar team achieved a similar quality as the reference translation. The quality of both translations were usually good, but we found some differences in some situations in which we preferred one translation over the other. For instance, in cases such as “out of 100” instead of “of 100”. Further, in one particular sentence, “rule out” was used as a translation for “discutir”, while the other used “discuss”. In many situations, we preferred translations that placed the verbs at the beginning of the sentence, such as in “We examined the absenteeism parameters...” instead of at the end, such as in “the parameters for granting time off work .... were analyzed”. Further, we find that the use of a specific and more suitable terms, such as “absenteeism”, “productivity”, and “control” are preferred to a longer or informal expression, such as “granting time off work”, “being productive” and “combat”, respectively.

en2es This year the overall quality of the translations was mixed. Both SRT and reference translations were of very good quality, and SRT was in many occasions indistinguishable to reference translations in the manual evaluation when it came to quality. However, the Huawei-BabelTar system had a mixed result with very good translations and translations of doubtful quality that clearly affected the fluency and readability of the output.

Capitalization and word separation were the main issues encountered when evaluating Huawei-BabelTar's output at a sentence level e.g. "La pandemia de covid-19 ofreció a la humanidad un portal a través del cual podemos romper con el pasado e imaginar nuestromundo de nuevo."

As in past years, the translation of acronyms and out-of-dictionary terminology remains a challenge for MT systems, Huwaei-BabelTar being a perfect example of such issues: "Describimos el caso de una cirrosis descompensada que desarrolló hpp y se resolvió con trasplante hepático, permaneciendo asintomática tras diez años de seguimiento."

When dealing with long named entities, word order remained a challenge for both SRT and Huawei-BabelTar, as in the following example where the numbers relate to the acronym "MMPs", and not to the noun "haplotypes":

(9) **Source**: To evaluate MMPs 7, 8, 12, and 13 haplotypes and their association with CRC.

Reference: Evaluar haplotipos de las MMP 7, 8, 12, y 13 y su asociación con CCR.

Huawei-BabelTar: Evaluar los haplotipos 7,8, 12 y 13 (incorrect word order and word separation) delmmp (word separation and capitalization) y su asociación con el ccr (capitalization of acronyms).

SRT: Evaluar los haplotipos 7, 8, 12 y 13 (incorrect word order) de MMP y su asociación con el CCR.

The reference translations were of very high quality overall, creating readable and fluent outputs at the level of sentences and abstracts. The main thing that differentiated reference translations from machine translations was the fact that they were less literal and followed writing conventions in Spanish for the domain, such as the use of passive reflexive tense which is more common in Spanish medical and scientific writings. However, the reference translation omitted relevant information or added implicit information from the text, which affect the overall quality of those translations when compared with the MT systems.

es2en This year the overall quality of the translations was mixed, with some very good quality translations coming from the MT systems (which made them nearly indistinguishable from the reference translations) to poorly written translations (including reference translations). Such is the case as well for the source text, which included some very high quality abstracts and also some poorly written abstractsn which contained grammatical errors such as lack of capitalization, wrong punctuation or word separation as in the following example:

(10) estudio observacional, relacional, transversal, en 185 derechohabientes de una unidad de medicina familiar del 15 de junio al 15 de agosto de 2020

This affected the quality of the output of both MT systems, Huwaei-BabelTar and SRT, which closely followed the source text:

(11) **Huwaei-BabelTar**: observational, relational, cross-sectional study in 185 beneficiaries of a family medicine unit from June 15 to August 15, 2020.

SRT: observational, relational, cross-sectional study in 185 beneficiaries of a family medicine unit from June 15 to August 15, 2020.

On the other hand, SRT proved to be more robst than Huwaei-BabelTar and the reference translation, and was able to deal with poor source text much more consistently such as in the example:

(12) **Source**: Existen múltiples causas delesiones ureterales, siendo la principal yatrogénica.

SRT: There are multiple causes of uretral injuries, the main one being iatrogenic.

Word order in longer sentences still remains a challenge for MT systems, which do not always correctly identify adverbs modifying long named entities as seen the following example, where "muchos" modifies the noun "biomarcadores":

(13) **Source** : La expansión y el descubrimiento de nuevas posibilidades de diagnóstico para el uso de muchos biomarcadores de enfermedades cardiovasculares (ECV), incluidas las isoformas de troponina cardioespecíficas (cTnI, cTnT), se debe a la mejora de los métodos de laboratorio para su determinación.

Huawei-BabelTar: The expansion and discovery of new diagnostic possibilities for the use of many cardiovascular disease (CVD) biomarkers, including cardio-specific troponin isoforms (cTnI, cTnT), is due to improved laboratory methods for their determination.

> **SRT**: The expansion and discovery of new diagnostic possibilities for the use of many cardiovascular disease (CVD) biomarkers, including cardio-specific troponin isoforms (cTnI, cTnT), is due to improved laboratory methods for their determination.

Both SRT and Huawei-BabelTar create sentences where "many" modifies "cardiovascular diseases", which changes the meaning of the translation in both cases.

However, the reference translations also had a mixed quality when compared to the MT systems, and presented issues such as poor capitalization or incorrect word separation, as seen in the following example: "There are many causesof ureteral injury being the main one iatrogenic".

Unlike previous years, SRT performed best in the three-way manual evaluation, coming close to the reference translation, due to the references' varied quality.

en2de Similarly to the last few years, the quality of the translations into German was very high. Both participants provided mostly convincing translations - partially including slight restructurings of the sentences. However, although the Huawei-BabelTar team performed lower in comparison to Huawei-TSC, the translations were in most cases not necessarily of lower quality. Instead, the Huawei-BabelTar system made two crucial errors, namely a) translations tend to ignore the capitalization of some German words, as well as b) single words were sometimes written together (without whitespace). Without those two error patterns, the quality of both translation systems would be closer to each other. Sometimes the systems used literal translatations, which impacted the quality of the translated text. For instance, "real-data" was translated into "reale Daten" (instead of "Daten aus der Praxis") or "essential" was translated into "essentiell" instead of "unerlässlich".

en2zh The translation quality this year was high. Unlike last year where few sentences were so awkwardly translated that a reader could hardly guess the original meaning, there were essentially no such sentences this year.

The biggest factor that reduced translation quality was the treatment of biomedical terms. This phenomenon came in two categories. The first category was straightforward, where the correct Chinese term was imprecise or downright wrong. For instance, *poor outcomes* of medical treatments was imprecisely translated as 不良结局 (poor end results), when the precise Chinese medical term was 不良预后 . In another example, *Rights-based Approaches (RBAs)* was translated as 基于权利的方法（非洲区域局） in which the full name of the term was correctly translated, but the abbreviation in brackets was incorrectly translated as *Regional Bureau Africa*.

The second category is more subtle, where the translated Chinese term was correct, but the presence of the original English term (or lack thereof) impacted readability. As an example, *Diabetic Retinopathy (DR)* was ideally translated as 糖尿病视网膜病变 *(Diabetic Retinopathy, DR)*, where the Chinese term, the English term, as well as the English abbreviation in the source text were all present. Another translation omitted the full English term, yielding 糖尿病视网膜病变 *(DR)*, which was still easily understandable. In another case, however, *healthcare workers (HCWs)* was translated to 医护人员 *(HCW)*. Here, the abbreviation was translated in singular form, conflicting with the plural form in the source text.

An interesting observation was that conventional, typical wording and punctuation in the translation significantly improved its quality. As a simple example, *experts disagreed* was translated by one system as 专家意见有分歧 (expert opinions differ) and by another as 专家们持不同看法 (experts have different opinions) . Both translations conveyed the same information, but the first translation was much more typical – refined even, as one would expect in a scientific publication. In terms of punctuation, the 、 is unique to the Chinese language when listing items. Hence when given *three overall reactions (positive, negative, and ambivalent)*, the translation 三种总体反应（积极、消极和矛盾） (note the punctuation between the first and second items) read much more naturally than 三种总体反应（积极，消极和矛盾） (exactly the same text, but a comma was used instead). In these cases, the less typical writing style was strictly speaking not wrong, but immediately hinted at the possibility that the text was not written by a native speaker.

Finally, the conversion between Western and Chinese number systems remained a challenge for some systems. The amount *598.851 billion yuan* referred to a billion as 10^9. The closest Chinese word to *billion* is 亿 , which is one order of

magnitude smaller at 10^8. This particular amount (598,851,000,000) was incorrectly converted to *598851 亿元* (59,885,100,000,000) by one system, and correctly though confusingly converted to *558851 m 元* (558,851 million) by another.

zh2en Continuing the trend from the previous two years, the translations this year are again of high quality. Nevertheless, a few common types of error still provide room for improvement.

Presumably, when a technical or medical term is missing from the system's dictionary, the individual Chinese characters in the term are translated literally. For instance, 清零 (zero-COVID policy) was translated by multiple systems as *zeroing*, which, despite the context of a COVID-related abstract, was hardly guessable. In another instance, 增强现实 (augmented reality) is arguably a technical term outside of the biomedical domain, but was still successfully translated as *augmented reality* by most systems and only one system produced *augmented real-world*.

In other cases, when a Chinese word has a general, non-biomedical meaning as well as a biomedical one, a system might incorrectly opt for the biomedical meaning. 服务阵地不断萎缩 (the continuous shrinking/decline of the service locations) is an example, where 萎缩 should be given the general translation of *decline* instead of the biomedical translation of *atrophy*.

When a translation overly preserves the fidelity of the source phrase, the resulting translation can be awkward. Take 在明确针刺可调节神经、血管这一共识的基础上 as an example. A more readable and thus preferable translation was *based on the consensus that acupuncture can regulate nerves and blood vessels*, even though a word-for-word translation would produce *on the basis of the consensus that acupuncture can regulate nerves and blood vessels* instead.

Similar to en2zh translations, numerical values also proved challenging for some systems in zh2en. 4.26 万 (one 万 is 10,000) is equivalent to 42,600, but the systems translated that variously to *4.26,000* or even *426 million*.

6.3 Targeted evaluation in clinical cases

This year, special attention was given to the evaluation of translations submitted by systems for the clinical case reports, from English into French.

The manual evaluation focused on the criteria that were used to select the clinical case: (1) acronyms; (2) numeric values including lab values; and (3) clinical correctness. Examples 14 and 16 illustrate erroneous translations produced by automatic systems while example 15 illustrates a case of an untranslated value. In the examples correct translations are shown in black font while incorrect ones appear in red fond. An asterisk indicates ungrammatical segments. Passages underlined within the same example block mark text that should carry the same meaning across statements.

(14) **en:** screening test for SARS-CoV-2
fr$_1$: dépistage systématique du SARS-CoV-2
fr$_2$: *dépistage systématique du CoV-2 du SARS

(15) **en:** the platelet count was 113 × 10E9/L
fr$_1$: les plaquettes sont à 113 000/mm^3
fr$_2$: numération plaquettaire de 113 10E09/L

(16) **en:** General examination revealed a wasted man
fr$_1$: L'examen clinique objective une dénutrition
fr$_2$: L'examen général a révélé un homme obèse

Specifically, the translations were annotated using BRAT[11] and aimed to assess the systems' performance on the specific aims. Annotations were produced independently by one annotator with formal translation training (AN) and one clinician (CG). For annotations on the full-text case descriptions, inter-annotator agreement on entities was high overall (above 0.75 F-measure) for "values" and "acronyms" and lower for "errors" (above 0.35 F-measure), mainly due to the identification of more errors by the clinician, which was expected. Inter-annotator agreement on attributes was medium overall (above 0.55 F-measure) mainly due to disagreements on "unclear" and "erroneous" translations, while agreement was much higher for "correct" and "untranslated" cases.

While the "correct" translation category was the most prevalent for all systems for values and acronyms, it can be noted that SPECTRANS produced more "Untranslated" occurrences. Overall, Huawei-BabelTar produced more "Errors" than the other two systems.

This analysis suggests that, in spite of high BLEU scores, the automatic translations can contain serious translation errors (e.g. Example 16)

[11] Brat Rapid Annotation Tool https://brat.nlplab.org/(Stenetorp et al., 2012)

and information that is not directly actionable for clinicians (e.g. Example 15).

7 Conclusions

We presented an overview of this year's edition of the WMT Biomedical Task. We released test sets for seven language pairs, and addressed a variety of textual sources, such as scientific abstracts, clinical case reports, and terminologies. We had a record number of participating teams and of submissions. All submissions were automatically evaluated in terms of BLEU scores, with respect to reference translations, whenever available. We also manually evaluated a selection of the submissions, and similar to previous years, the translations from some teams achieved a similar quality to the reference translations.

Limitations

The scope of the biomedical task has been growing over the years. While each new edition builds on the experience of the previous one, the scale of operations implies a number of limitations from operational and theoretic perspectives. One major limitation is the comparison between translation approaches used by the teams. The information we collect through the participant survey attempts to document the material and methods used by the participants' systems. However, it can be noted that only a subset of teams do supply details of their systems. Furthermore, some descriptions such as the training corpus size or content could be clarified. A closed task, where all participants are limited to using specific training material, could help improve comparability but would require additional work from participants and organizers. Another limitation is the imbalance between language pairs, which attracts different levels of effort from both participants and organizers.

MT can be computationally intensive and the environmental impact of experiments should be measured. While no measure of impact was conducted this year, we included this aspect in the participant survey, which included a list of tools that can be used to measure impact. A future growth direction to increase awareness of impact can be to ask participants to supply a measure of CO2 impact along with their results.

Ethics Statement

This task mainly focuses on translation using the MEDLINE corpus, which is openly available for research. The test corpora used in the task were selected based on publication date and linguistic criteria. Any imbalance regarding the demographics of populations represented in the corpus is involuntary.

The intended use of this task is to contribute to the evaluation and training of MT systems in the biomedical domain. We do not recommend the use of MT without expert validation in a medical context, as machine translated text could contain errors impacting patients' health outcomes.

Acknowledgements

The ClinSpen sub-task was supported by the Spanish Government's Encargo of Plan TL (SEAD) to BSC and BIOMATDB project of European Union's Horizon Europe Coordination & Support Action under Grant Agreement No 101058779, as well as AI4PROFHEALTH project (PID2020-119266RA-I00).

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.

References

Nicolas Ballier, Jean-Baptiste Yunès, Guillaume Wisniewski, Lichao Zhu, and Maria Zimina. 2022. The SPECTRANS System Description for the WMT22 Biomedical Task. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of nlp methods: a survey and analysis of existing tools. In *EMNLP, Workshop SustaiNLP*.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor,

and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Yoonjung Choi, Jiho Shin, Yonghyun Ryu, and Sangha Kim. 2022. SRT's Neural Machine Translation System for WMT22 Biomedical Translation Task. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

E Ezhergina, M Fedorova, V Malykh, and D Petrova. 2022. Findings of biomedical Russian-English MT competition. In *AINL 2022 Proceedings*.

Christian Federmann. 2010. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 1731–1734, Valletta, Malta.

Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 Biomedical Translation Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark.

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.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ernan Li, Fandong Meng, and Jie Zhou. 2022. Summer: WeChat Neural Machine Translation Systems for the WMT22 Biomedical Translation Task. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Boxiang Liu and Liang Huang. 2021. Paramed: a parallel corpus for english–chinese translation in the biomedical domain. *BMC Medical Informatics and Decision Making*, 21(1):1–11.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 Biomedical Translation Shared Task: Evaluation on MEDLINE test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339. Association for Computational Linguistics.

Aurélie Névéol, Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2018. Parallel Corpora for the Biomedical Domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera,

Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Weixuan Wang, Xupeng Meng, Suqing Yan, Ye TIAN, and Wei Peng. 2022. Huawei BabelTar NMT at WMT22 Biomedical Translation Task: How we further improve domain-specific NMT. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, Yuanchang Luo, Yuhao Xie, Miaomiao Ma, Ting Zhu, Lizhi Lei, Song Peng, Hao Yang, and Ying Qin. 2022. HW-TSC Translation Systems for the WMT22 Biomedical Translation Task. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.

Huanran Zheng, Wei Zhu, and Xiaoling Wang. 2022. ECNU-MT's WMT22 Biomedical Translation Task Submission. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

A Mapping of runs in OCELoT

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh
aoligei	run1	-	-	-	-	-	-	24
	run1	-	-	-	-	-	-	25*
Dtranx	run1	360*	283*	277	304	305	352*	355*
	run2	-	291	416	418	421	469	379
	run3	-	-	448*	449*	450*	-	-
Huawei-TSC	run1	251	-	517	-	-	659	249
	run2	395	-	534	-	-	663	396
	run3	480	-	645	-	-	671*	478
	run4	520	-	774*	-	-	-	536
	run5	832	-	-	-	-	-	636
	run6	837*	-	-	-	-	-	778
Lan-BridgeMT	run1	115	-	-	-	-	113*	9
	run2	201*	-	-	-	-	198	177
	run3	-	-	-	-	-	-	202
	run4	-	-	-	-	-	-	388*
njupt-mtt	run1	140	-	124	-	-	142	88
	run2	-	-	128	-	-	155	90
	run3	-	-	579	-	-	163	93
SPECTRANS	run1	-	-	398	-	-	-	-
	run2	-	-	460*	-	-	-	-
	run3	-	-	484	-	-	-	-
	run4	-	-	486	-	-	-	-
ustc-mt	run1	-	-	312	-	-	345	722
	run2	-	-	314	-	-	369	764
	run3	-	-	542	-	-	-	-
	run4	-	-	543	-	-	-	-
	run5	-	-	544	-	-	-	-
	run6	-	-	569	-	-	-	-

Table 23: Mapping of the MEDLINE runs to the submission ids in OCELoT Biomedical Task, from English. An asterisk * indicates the primary run.

Teams	Runs	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en
aoligei	run1	-	-	-	-	-	-	17
	run2	-	-	-	-	-	-	20
	run3	-	-	-	-	-	-	21*
	run4	-	-	-	-	-	-	23
DTranx	run1	357	303*	306	307	308	353	356
	run2	472*	-	422	423	425	470*	471*
	run3	473	-	451*	452*	453*	-	-
Huawei-TSC	run1	252	-	646	-	-	596	250
	run2	411	-	732	-	-	597	397*
	run3	481	-	750	-	-	600	479
	run4	537*	-	-	-	-	601*	523
	run5	-	-	-	-	-	-	638
	run6	-	-	-	-	-	-	781
Lan-BridgeMT	run1	104	-	-	-	-	105*	19
	run2	200*	-	-	-	-	199	203
	run3	-	-	-	-	-	-	220
	run4	-	-	-	-	-	-	221
	run5	-	-	-	-	-	-	387*
njupt-mtt	run1	139	-	125	-	-	145	89
	run2	-	-	129	-	-	156	95
	run3	-	-	580	-	-	-	-
SPECTRANS	run1	-	-	399	-	-	-	-
	run2	-	-	462*	-	-	-	-
	run3	-	-	487	-	-	-	-
	run4	-	-	492	-	-	-	-
ustc-mt	run1	-	-	316	-	-	346	724
	run2	-	-	317	-	-	371	-
	run3	-	-	565	-	-	-	-
	run4	-	-	567	-	-	-	-
	run5	-	-	568	-	-	-	-
szdx	-	-	-	-	-	-	-	97

Table 24: Mapping of the runs to the submission ids in OCELoT Biomedical task, into English. An asterisk * indicates the primary run.

Teams	Runs	en2de	en2ru	en2zh
AISP-SJTU	run1	-	-	31
	run2	-	-	611
ALMAnaCH-Inria	run1	-	381	-
	run2	-	711	-
aoligei	run1	-	-	26
	run2	-	-	27
bhcs-mt	run1	-	-	43
	run2	-	-	44
	run3	-	-	170
	run4	-	-	172
DLUT	run1	-	-	430
	run2	-	-	649
	run3	-	-	651
	run4	-	-	721
Dtranx	run1	319	329	333
	run2	325	461	354
	run3	765	-	-
eTranslation	run1	-	337	-
	run2	-	339	-
	run3	-	341	-
GTCOM	run1	-	-	521
	run2	-	-	733
	run3	-	-	853
HuaweiTSC	run1	-	-	236
	run2	-	-	465
	run3	-	-	476
	run4	-	-	557
	run5	-	-	575
	run6	-	-	630
	run7	-	-	776
JDExploreAcademy.Vega-MT	run1	507	509	59
	run2	843	690	98
	run3	-	-	102
	run4	-	-	833
	run5	-	-	652
	run6	-	-	706
	run7	-	-	834
KwaiMT	run1	-	-	794
	run2	-	-	797
	run3	-	-	799
Lan-Bridge	run1	114	112	12
	run2	191	197	162
	run3	393	409	175
	run4	549	556	714
LanguageX	run1	-	-	150
	run2	-	-	692
	run3	-	-	701
	run4	-	-	716
Manifold	run1	-	-	28
	run2	-	-	136
	run3	-	-	231
	run4	-	-	336
	run5	-	-	440
	run6	-	-	604
	run7	-	-	820
MeteorMan	run1	-	-	230
neunlplab	run1	-	-	14
	run2	-	-	67
	run3	-	-	570
	run4	-	-	760
	run5	-	-	798
	run6	-	-	847

Table 25: Mapping of the runs to the submission ids in OCELoT General Task, from English (part 1/2).

Teams	Runs	en2de	en2ru	en2zh
njupt-mtt	run1	-	137	85
	run2	-	147	92
	run3	-	213	144
	run4	-	243	211
	run5	-	-	214
	run6	-	-	216
	run7	-	-	232
ONLINE-A	run1	901	912	914
	run2	-	911	-
ONLINE-B	run1	920	930	931
	run2	-	-	932
Online-G	run1	865	876	878
ONLINE-W	run1	954	966	968
	run2	959	-	-
ONLINE-Y	run1	939	949	951
	run2	973	983	985
OpenNMT	run1	207	-	-
	run2	210	-	-
	run3	321	-	-
	run4	493	-	-
	run5	746	-	-
PROMT	run1	68	42	-
	run2	334	71	-
	run3	694	72	-
	run4	-	73	-
	run5	-	804	-
SRPOL	run1	-	157	-
	run2	-	160	-
	run3	-	265	-
	run4	-	496	-
	run5	-	497	-
	run6	-	501	-
super_star	run1	-	-	228
	run2	-	-	229
szdx	run1	-	-	119
	run2	-	-	338
	run3	-	-	436
	run4	-	-	438
	run5	-	-	439
	run6	-	-	441
	run7	-	-	442
taicangshaxigaozhong	run1	-	-	788
	run2	-	-	811
ustc-mt	run1	-	-	276
	run2	-	-	279
	run3	-	-	281
	run4	-	-	293
	run5	-	-	328
	run6	-	-	373
V2ray	run1	-	-	47

Table 26: Mapping of the runs to the submission ids in OCELoT General Task, from English (part 2/2).

Teams	Runs	de2en	ru2en	zh2en
AISP-SJTU	-	-	-	648
ALMAnaCH-Inria	run1	-	382	-
	run2	-	710	-
aoligei	run1	-	-	11
	run2	-	-	146
	run3	-	-	151
	run4	-	-	154
	run5	-	-	295
bhcs-mt	run1	-	-	45
	run2	-	-	171
	run3	-	-	173
	run4	-	-	737
	run5	-	-	810
bymt	run1	-	-	294
Dtranx	run1	315	343	349
	run2	429	463	468
	run3	456	-	-
DLUT	run1	-	-	432
	run2	-	-	653
	run3	-	-	654
HuaweiTSC	run1	-	-	245
	run2	-	-	467
	run3	-	-	571
	run4	-	-	626
JDExploreAcademy.Vega-MT	run1	508	510	58
	run2	809	769	99
	run3	-	844	101
	run4	-	-	656
	run5	-	-	658
	run6	-	-	708
	run7	-	-	736
KwaiMT	run1	-	-	415
	run2	-	-	790
	run3	-	-	792
Lan-Bridge	run1	103	86	10
	run2	188	187	222
	run3	587	589	223
	run4	-	-	386
LanguageX	run1	-	-	168
	run2	-	-	218
	run3	-	-	219
	run4	-	-	400
	run5	-	-	412
	run6	-	-	417
Liaoning University	run1	-	-	152
	run2	-	-	498
	run3	-	-	830
LT22	run1	605	-	-
	run2	608	-	-
	run3	612	-	-
	run4	614	-	-
	run5	617	-	-
neunlplab	run1	-	-	14
	run2	-	-	67
	run3	-	-	570
	run4	-	-	760
	run5	-	-	798
	run6	-	-	847

Table 27: Mapping of the runs to the submission ids in OCELoT General Task, into English (part 1/2).

Teams	Runs	de2en	ru2en	zh2en
njupt-mtt	run1	-	138	87
	run2	-	153	143
	run3	-	254	212
	run4	-	-	215
	run5	-	-	217
	run6	-	-	237
	run7	-	-	244
ONLINE-A	run1	903	913	915
ONLINE-B	run1	923	934	935
Online-G	run1	868	861	879
ONLINE-W	run1	956	967	969
	run2	961	-	-
ONLINE-Y	run1	941	950	952
	run2	975	984	986
pingan_mt	-	-	-	494
PROMT	run1	29	70	-
	run2	796	-	-
SRPOL	run1	-	272	-
	run2	-	359	-
	run3	-	361	-
	run4	-	661	-
	run5	-	664	-
	run6	-	666	-
	run7	-	697	-
star	run1	-	-	296
	run2	-	-	297
	run3	-	-	602
	run4	-	-	665
super_star	run1	-	-	159
	run2	-	-	166
	run3	-	-	165
	run4	-	-	167
	run5	-	-	227
	run6	-	-	242
	run2	-	-	166
	run3	-	-	165
	run4	-	-	167
	run5	-	-	227
	run6	-	-	242
szdx	run1	-	-	100
	run2	-	-	123
	run3	-	-	134
	run4	-	-	599
	run5	-	-	631
	run6	-	-	634
	run7	-	-	635
taicangshaxigaozhong	run1	-	-	618
	run2	-	-	640
	run3	-	-	791
	run4	-	-	813
ustc-mt	run1	-	-	280
	run2	-	-	282
	run3	-	-	292
	run4	-	-	327
	run5	-	-	477
V2ray	run1	-	-	48

Table 28: Mapping of the runs to the submission ids in OCELoT General Task, into English (part 2/2).