Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Iñigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, et al.

▶ **To cite this version:**

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, et al.. Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set. EMNLP 2021 - Sixth Conference on Machine Translation., Nov 2021, Punta Cana, Dominican Republic. hal-03435096

HAL Id: hal-03435096
https://hal.science/hal-03435096v1

Submitted on 18 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set

Lana Yeganova[1*] **Dina Wiemann**[2] **Mariana Neves**[3] **Federica Vezzani**[4]
Amy Siu[5] **Iñigo Jauregi Unanue**[6] **Maite Oronoz**[7] **Nancy Mah**[8]
Aurélie Névéol[9] **David Martinez**[10,11] **Rachel Bawden**[12] **Giorgio Maria Di Nunzio**[13]
Roland Roller[14] **Philippe Thomas**[14] **Cristian Grozea**[15] **Olatz Perez de Viñaspre**[7]
Maika Vicente Navarro[16] **Antonio Jimeno Yepes**[10]

[1]NCBI/NLM/NIH, Bethesda, USA
[2]Novartis AG, Basel, Switzerland
[3]German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[4]Dept. of Linguistic and Literary Studies University of Padua, Italy
[5]Berliner Hochschule für Technik, Germany
[6]University of Technology Sydney, Sydney, Australia
[7]IXA NLP Group, University of the Basque Country, Donostia, Spain
[8]Fraunhofer Institute for Biomedical Engineering (IBMT), Berlin, Germany
[9]LISN, CNRS, Université Paris-Saclay, Orsay, France
[10]University of Melbourne, Australia
[11]Doctor Evidence, Santa Monica, CA, USA
[12]Inria, France
[13]Dept. of Information Engineering, University of Padua, Italy
[14]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[15]Fraunhofer Institute FOKUS, Berlin, Germany
[16]Maika Spanish Translator, Melbourne, Australia

Abstract

In the sixth edition of the WMT Biomedical Task, we addressed a total of eight language pairs, namely English/German, English/French, English/Spanish, English/Portuguese, English/Chinese, English/Russian, English/Italian, and English/Basque. Further, our tests were composed of three types of textual test sets. New to this year, we released a test set of summaries of animal experiments, in addition to the test sets of scientific abstracts and terminologies. We received a total of 107 submissions from 15 teams from 6 countries.

1 Introduction

Machine translation (MT) is the automatic translation of textual resources from one language to another. It is an important component in many applications and natural language processing (NLP) pipelines in the clinical and biomedical domains. On the one hand, some resources, such as specific biomedical terminologies, are only available for a limited number of languages. English is especially well covered in the Unified Medical Language System (UMLS) (Lindberg et al., 1993) while other languages are not (Wilde, 2021). On the other hand, there are many publications written in languages other than English and are therefore inaccessible to researchers who cannot read those languages.

This context has been the overarching goal for the organization of the WMT Biomedical task. The first edition took place in 2016 and addressed scientific abstracts for English/French (both directions), English/Spanish (both directions), and En-

[*] The organization of the biomedical task is complex and relies on varied essential contributions from many individuals. Authors are listed randomly because we could not do justice to the contributors using a single ranking. We would like to acknowledge MN for dataset preparation and general task organization, CG for creating baselines, AN for compiling information on participants methods, AJY for conducting the automatic evaluation, LY, DW, MN, FV, AS, AN, GMDN, RR, PT, MVN, AJY for evaluating the alignment of the test sets, and LY, DW, MN, FV, AS, MO, NM, AN, RB, GMDN, RR, PT, MVN, AJY for conducting the manual evaluation. All authors approved the final version of the manuscript. E-mail for contact: mariana.lara-neves@bfr.bund.de

glish/Portuguese (both directions) (Bojar et al., 2016). The subsequent shared task included six new language pairs, namely, English into Czech, English into German, English into Hungarian, English into Polish, English into Romanian, and English into Swedish, in addition to a new type of document, viz., health information texts (Jimeno Yepes et al., 2017). In 2018, we started using MEDLINE® as the source for our scientific abstracts and addressed a new language pair, namely English/Chinese (both directions), in addition to some of the languages already considered in the previous year (Neves et al., 2018). In the subsequent year, we introduced the translation of biomedical terminologies (from English into Spanish), in addition to the MEDLINE abstracts for the five language pairs from the 2018 task (Bawden et al., 2019). In 2020, we added three new language pairs, namely English/Russian (both directions), English/Italian (both directions), and English into Basque (en2eu) (Bawden et al., 2020).

For this year's shared task[1], we address the same eight language pairs as last year (Bawden et al., 2020) on the same translations tasks (scientific abstracts and terminologies). The main novel feature this year is a new test set composed of summaries of planned animal experiments to be translated from German into English. The list below summarizes the language pairs addressed this year:

- English to Basque (en2eu)
- English to Chinese (en2zh) and Chinese to English (zh2en)
- English to French (en2fr) and French to English (fr2en)
- English to German (en2de) and German to English (de2en)
- English to Italian (en2it) and Italian to English (it2en)
- English to Portuguese (en2pt) and Portuguese to English (pt2en)
- English to Russian (en2ru) and Russian to English (ru2en)
- English to Spanish (en2es) and Spanish to English (es2en)

Finally, we highlight the new aspect that we introduced in the 2021 edition of our shared task, namely, a novel test set for the automatic translation of summaries of animal experiments from German into English (see Section 2.4).

2 Training and test data

No additional training data was released for any of the language pairs, with the exception of en2eu, where we provide last year's test set as new training data for abstracts and terminology. As for the tests sets, we released test sets for scientific abstracts, terminologies, and summaries of animal experiments as follows:

- Scientific abstracts:
  - English to Basque
  - Chinese/English (both directions)
  - French/English (both directions)
  - German/English (both directions)
  - Italian/English (both directions)
  - Portuguese/English (both directions)
  - Russian/English (both directions)
  - Spanish/English (both directions)
- Terms from biomedical terminologies:
  - English to Basque
- Summaries of animal experiments:
  - German to English

Table 1 shows the number of documents, sentences and terms (if applicable) for each test set. In this section, we give details on the construction of the test sets.

2.1 MEDLINE test sets

Similar to previous years, we retrieved recent MEDLINE abstracts that were available in both English and one of the seven other languages we evaluate on (namely Chinese, French, German, Italian, Portuguese, Russian, and Spanish). The abstracts in both languages were processed as follows:

- language detection with the Python `langdetect` library;[2]
- sentence splitting using the Python `syntok` library;[3]

[1] http://www.statmt.org/wmt21/biomedical-translation-task.html

[2] https://pypi.org/project/langdetect/

[3] https://github.com/fnl/syntok

Language pairs	Abstracts		Terminology	Summaries	
	Documents	Sentences	Terms	Documents	Sentences
en2eu	76	450	2,736	-	-
de2en	50	480/481	-	30	648
en2de	50	516/501	-	-	-
es2en	50	445/444	-	-	-
en2es	50	486/501	-		
fr2en	50	365/351	-	-	-
en2fr	50	384/394	-		
it2en	43	432/407	-	-	-
en2it	44	460/448	-		
pt2en	50	468/484	-	-	-
en2pt	50	494/486	-		
ru2en	50	428/436	-	-	-
en2ru	50	354/373	-		
zh2en	50	341/393	-	-	-
en2zh	50	425/375	-		

Table 1: Number of documents, sentences and terms in the test sets released for this shared task. Some abstracts had to be removed from the it2en and en2it during the evaluation phase.

- sentence alignment using the `GMA` tool[4] for all language pairs except for English/Chinese, for which the `Champollion` tool[5] was used;
- random retrieval of 100 abstracts for each language pair;
- and manual validation of the selected abstracts using the "quality checking" task in the Appraise tool (Federmann, 2010), of which the results are shown in Table 2.

Table 2 shows that the highest quality was obtained for the zh/en test sets, with up to 94.5% perfectly aligned sentences. This is actually not a surprise, since these were the only test sets where an expert manually discarded abstracts that are clearly non-parallel, e.g. when the entire English abstract corresponds to only the first half of the Chinese abstract. A high quality of over 80% was also obtained for four language pairs, namely pt/en (90.4%), es/en (88.4%), fr/en (86.0%), and it/en (80.3%).

For en/fr, the automatic alignment was manually reviewed. In this process, the overall corpus size increased from 630 sentences to 775 sentences, mainly through the addition of article titles that had not been collected in English, and did not have any equivalent in French. In terms of alignment quality, it is important to note that the problematic categories *Source>Target* and *Target>Source* are significantly reduced in the revised corpus. The de/en test set obtained a slightly lower quality of 77.7%, while only 54.2% of ru/en sentences were perfectly aligned. Similar to previous years, the automatic evaluation was carried out for all sentences as well as only for the perfectly aligned (hereafter referred to as "OK") ones.

2.2 Basque abstracts

As we mentioned in (Bawden et al., 2020), the presence of Basque in MEDLINE is almost non-existent. In this edition we have again used the abstracts from the journal Osagaiz[6] as part of the test set, but due to the low production of this journal written in Basque, we have added abstracts from the journal *Gaceta Médica de Bilbao*[7], which contains abstracts written in Spanish, English, and Basque. From the 76 documents and 450 sentences mentioned in table 1, 18 documents and 119 sentences are from the Osagaiz journal, and 50 documents and 331 sentences from *Gaceta Médica de Bilbao*. The sentences were manually aligned by human annotators.

2.3 Terminologies

In the WMT20 edition, on behalf of Osakidetza (Basque Public Health System), we released 27,900 terms of the Basque ICD-10-CM edition, 2,000 of

[4] https://nlp.cs.nyu.edu/GMA/
[5] http://champollion.sourceforge.net/
[6] http://www.osagaiz.eus
[7] http://www.gacetamedicabilbao.eus/index.php/gacetamedicabilbao

Language	OK	Source>Target	Target>Source	Overlap	No Align.	Total
de/en	710 (77.7%)	38 (4.2%)	40 (4.4%)	34 (3.7%)	92 (10.0%)	914
es/en	792 (88.4%)	55 (6.1%)	15 (1.7%)	7 (0.8%)	27 (3.0%)	896
fr/en	540 (85.7%)	68 (10.8%)	10 (1.6%)	1 (0.2%)	11 (1.7%)	630
fr/en §	666 (86.0%)	9 (1.2%)	1 (0.1%)	8 (1.0%)	91 (11.7%)	775
it/en	666 (80.3%)	51 (6.2%)	26 (3.1%)	13 (1.6%)	73 (8.8%)	829
pt/en	838 (90.4%)	54 (5.8%)	18 (2.0%)	15 (1.6%)	2 (0.2%)	927
ru/en	371 (54.2%)	79 (11.5%)	63 (9.2%)	25 (3.6%)	147 (21.5%)	685
zh/en	658 (94.5%)	16 (2.3%)	9 (1.3%)	1 (0.2%)	12 (1.7%)	696

Table 2: Statistics (number of sentences and percentages) of the quality of the automatic alignment for the MEDLINE test sets. For each language pair, the total number of sentences corresponds to the 100 documents that constitute the two test sets (one for each language direction). § Results after manual correction of sentence segmentation and/or alignment.

which were used for evaluation. This year, we updated some of the Basque translations for correctness and cohesion. The full set from last year was released for training and a new set of 2,736 terms was used as a test set.

2.4 Summaries of planned animal experiments

We released a test set of 30 summaries of planned animal experiments that were retrieved from the AnimalTestInfo database[8], which is maintained by the German Federal Institute for Risk Assessment (BfR). The summaries describe planned and approved animal experiments to be carried out in Germany, which are anonymously stored in this online database in a bid to improve transparency (Bert et al., 2017). The aim of considering these summaries in this shared task is to assess the quality of MT of these documents, which is relevant for a couple of projects currently being carried out in the BfR, such as mining for alternative methods to animal experiments. A previous larger training set and test set from this database has been previously used in another shared task for the assessment of the automatic assignment of ICD-10 codes (Neves et al., 2019). The summaries contain following information (see Figure 1 for an example):

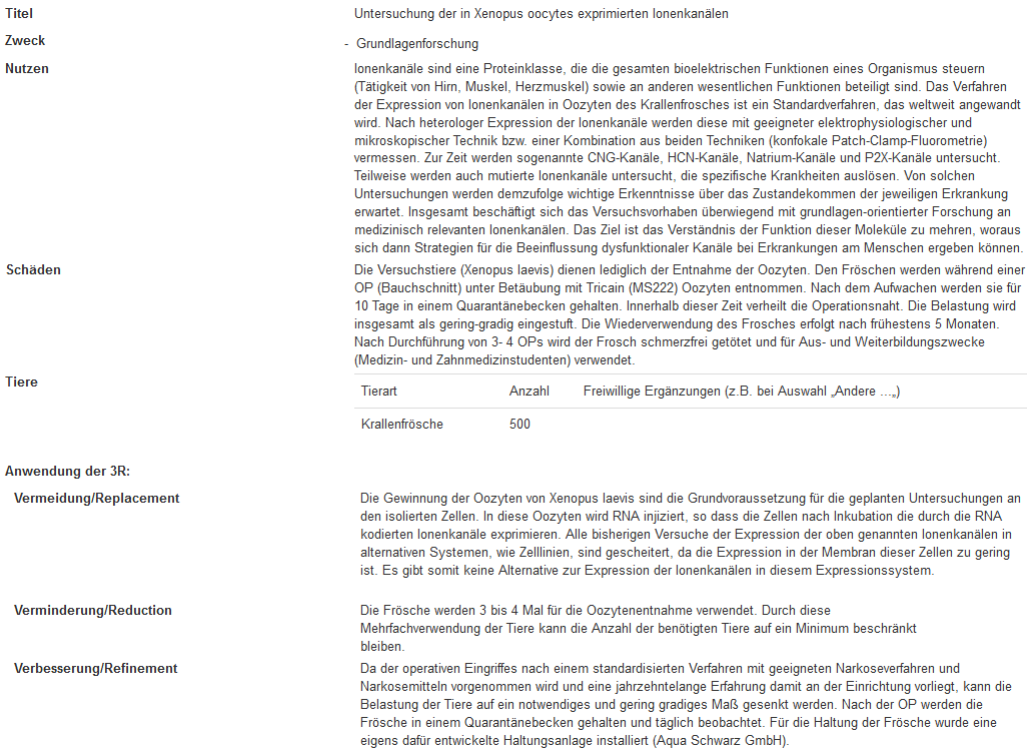
- title;
- aim of the study (e.g., basic research);
- benefits of the experiments;
- species and number of animals to be used;
- comments regarding the compliance to the so-called 3R principle (replacement, reduction, refinement of animal experiments).

The summaries were selected from the database in a way that addressed various animal species, and they were then manually translated by an English native speaker with a high knowledge of German. Before releasing the data, we converted the summaries into a format that is suitable for the WMT shared task.

3 Baselines

This year we had more choices for the baselines. As before, one option was to use our own models, trained with Marian NMT (Junczys-Dowmunt et al., 2018) on biomedical texts. A new option was to use pre-trained models, not specialized on biomedical texts. We used our own models as baselines for the following language directions: en2de, en2es, en2fr, en2pt, de2en, es2en, fr2en, pt2en. For en2zh, en2ru, en2it, en2eu, zh2en, ru2en, it2en, we used the pre-trained generic Marian NMT models available in the HuggingFace "Transformers" library.[9] An interesting question was whether the specialized models were still better than the newest out-of-the-box pretrained models. To this end, for en2de and en2fr we also tested the recent T5-large[10] model (Raffel et al., 2019). For en2fr, it outperformed our own model (trained on biomedical data) by almost 3 BLEU points, whereas for en2de the two systems were fairly comparable. The performance of the systems submitted starts from close to baseline for some language directions (e.g. for en2fr, en2es, de2en), whereas for other languages all systems were much better than the baseline (e.g. es2en, pt2en and especially ru2en).

[8] https://animaltestinfo.de/
[9] https://huggingface.co/Helsinki-NLP
[10] https://huggingface.co/t5-large



Titel Untersuchung der in Xenopus oocytes exprimierten Ionenkanälen

Zweck - Grundlagenforschung

Nutzen Ionenkanäle sind eine Proteinklasse, die die gesamten bioelektrischen Funktionen eines Organismus steuern (Tätigkeit von Hirn, Muskel, Herzmuskel) sowie an anderen wesentlichen Funktionen beteiligt sind. Das Verfahren der Expression von Ionenkanälen in Oozyten des Krallenfrosches ist ein Standardverfahren, das weltweit angewandt wird. Nach heterologer Expression der Ionenkanäle werden diese mit geeigneter elektrophysiologischer und mikroskopischer Technik bzw. einer Kombination aus beiden Techniken (konfokale Patch-Clamp-Fluorometrie) vermessen. Zur Zeit werden sogenannte CNG-Kanäle, HCN-Kanäle, Natrium-Kanäle und P2X-Kanäle untersucht. Teilweise werden auch mutierte Ionenkanäle untersucht, die spezifische Krankheiten auslösen. Von solchen Untersuchungen werden demzufolge wichtige Erkenntnisse über das Zustandekommen der jeweiligen Erkrankung erwartet. Insgesamt beschäftigt sich das Versuchsvorhaben überwiegend mit grundlagen-orientierter Forschung an medizinisch relevanten Ionenkanälen. Das Ziel ist das Verständnis der Funktion dieser Moleküle zu mehren, woraus sich dann Strategien für die Beeinflussung dysfunktionaler Kanäle bei Erkrankungen am Menschen ergeben können.

Schäden Die Versuchstiere (Xenopus laevis) dienen lediglich der Entnahme der Oozyten. Den Fröschen werden während einer OP (Bauchschnitt) unter Betäubung mit Tricain (MS222) Oozyten entnommen. Nach dem Aufwachen werden sie für 10 Tage in einem Quarantänebecken gehalten. Innerhalb dieser Zeit verheilt die Operationsnaht. Die Belastung wird insgesamt als gering-gradig eingestuft. Die Wiederverwendung des Frosches erfolgt nach frühestens 5 Monaten. Nach Durchführung von 3- 4 OPs wird der Frosch schmerzfrei getötet und für Aus- und Weiterbildungszwecke (Medizin- und Zahnmedizinstudenten) verwendet.

Tiere

Tierart	Anzahl	Freiwillige Ergänzungen (z.B. bei Auswahl „Andere ...„)
Krallenfrösche	500	

Anwendung der 3R:

Vermeidung/Replacement Die Gewinnung der Oozyten von Xenopus laevis sind die Grundvoraussetzung für die geplanten Untersuchungen an den isolierten Zellen. In diese Oozyten wird RNA injiziert, so dass die Zellen nach Inkubation die durch die RNA kodierten Ionenkanäle exprimieren. Alle bisherigen Versuche der Expression der oben genannten Ionenkanälen in alternativen Systemen, wie Zelllinien, sind gescheitert, da die Expression in der Membran dieser Zellen zu gering ist. Es gibt somit keine Alternative zur Expression der Ionenkanälen in diesem Expressionssystem.

Verminderung/Reduction Die Frösche werden 3 bis 4 Mal für die Oozytenentnahme verwendet. Durch diese Mehrfachverwendung der Tiere kann die Anzahl der benötigten Tiere auf ein Minimum beschränkt bleiben.

Verbesserung/Refinement Da der operativen Eingriffes nach einem standardisierten Verfahren mit geeigneten Narkoseverfahren und Narkosemitteln vorgenommen wird und eine jahrzehntelange Erfahrung damit an der Einrichtung vorliegt, kann die Belastung der Tiere auf ein notwendiges und gering gradiges Maß gesenkt werden. Nach der OP werden die Frösche in einem Quarantänebecken gehalten und täglich beobachtet. Für die Haltung der Frösche wurde eine eigens dafür entwickelte Haltungsanlage installiert (Aqua Schwarz GmbH).

Figure 1: Example of a summary for a planned animal experiment. Source: https://animaltestinfo.de/dsp_show_ntp.cfm?ntpID=19362

4 Teams and systems

This year, we received a total of 107 runs from 15 teams from the following countries: China (8), Spain (2), France (1), Japan (1), Pakistan (1), and USA (2). Table 3 presents the list of teams. We can note four returning teams: Huawei_AGI (most team members were part of the Huawei United team in 2020), LISN (LIMSI in 2020), nrpu-fjwu and TMT.

Table 4 presents an overview of the runs submitted by each team for language directions translating *from* English. Table 5 presents an overview of the runs submitted by each team for language directions translating *into* English.

We did not receive any submission for en2pt, even though we did receive submissions from one team for the opposite direction of this language pair, i.e., pt2en. Unfortunately, we did not receive any submission for the new test set that we released this year, i.e., for the summaries of planned animal experiments. Nevertheless, the test set (including the reference translation) is available for the research community for further experiments.

Similarly to the WMT 2020 biomedical task edition, we asked participants to fill in a survey with key information regarding the specific material and methods used in their self-identified primary runs that were used for manual evaluation. The survey comprised 14 questions covering the translation methods and corpora used.

On average, the time spent by participants to supply information for one language pair was 6 minutes and 35 seconds (Median: 3 minutes and 27 seconds). This is consistent with the 2020 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) and largely relied on existing implementations: 7 teams submitted runs using available libraries while 5 teams submitted runs using their own NMT implementations. Teams often used the same setup for a range of language pairs. Table 6 shows 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 corpus used for Chinese have been prepared by the teams but are not always available or described in details. We can also notice that the use or pre-processing of resources supplied by the task organizers can differ between teams as the size reported for seemingly similar data can differ sig-

Team ID	Institution
ECNU_PAHT	Pingan Health Tech / ECNU, China
FJDMATH (Martínez, 2021)	Fujitsu DMATH, Japan
Haozhiweizi	Shanghai Jiaotong University, China
Huawei_AGI (Wang et al., 2021a)	Huawei Technologies, China
Huawei_TSC (Yang et al., 2021)	Huawei Translation Service Center, China
JinDong	unknown, China
LISN	LISN, CNRS, France
MT Learner	Microsoft Research, China
NVIDIA NeMo (Subramanian et al., 2021)	NVIDIA, USA
nrpu-fjwu (Naz et al., 2021)	Fatima Jinnah Women University, Pakistan
talp_upc (Rafieian and Costa-Jussà, 2021)	Universitat Politècnica de Catalunya, Spain
TMT (Wang et al., 2021b)	Tencent AI Lab, China
Transperfect	Transperfect Translations, Spain
Volctrans	ByteDance, China
ZengHuiMT	FORYOR HEALTH, USA

Table 3: List of the participating teams.

Teams	en2eu	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh	Total
ECNU_PAHT	-	-	-	-	-	-	-	A3	3
FJDMATH	T2A2	-	-	-	-	-	-	-	4
Haozhiweizi	-	-	-	-	-	-	-	A1	1
Huawei_AGI	-	A3	-	A3	A3	-	-	A3	12
Huawei_TSC	-	A3	-	-	-	-	-	A3	6
JingDong	-	-	-	-	-	-	-	A1	1
LISN	-	-	-	A3	-	-	-	-	3
NVIDIA NeMo	-	-	-	-	-	-	A2	-	2
talp_upc	-	-	A2	-	-	-	-	-	2
TMT	-	A1	A1	A1	-	-	A1	-	4
Transperfect	-	-	A3	-	-	-	A2	A2	7
Volctrans	-	-	-	-	-	-	-	A3	3
ZengHuiMT	-	-	-	-	-	-	-	A1	1
Total	4	7	6	7	3	0	5	17	49

Table 4: Overview of the submissions from all teams and test sets translating from English. We identify submissions to the abstracts testsets with an "A" and to the terminology test set with a "T". The value next to the letter indicates the number of runs for the corresponding test set, language pair, and team.

Teams	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en	Total
ECNU_PAHT	-	-	-	-	-	-	A3	3
Haozhiweizi	-	-	-	-	-	-	A1	1
Huawei_AGI	A3	-	A3	A3	-	-	A3	12
Huawei_TSC	A3	-	-	-	-	-	A3	6
JingDong	-	-	-	-	-	-	A1	1
LISN	-	-	A3	-	-	-	-	3
MT Learner	-	-	-	A2	A2	A2	-	6
NVIDIA NeMo	-	-	-	-	-	A1	-	1
nrpu-fjwu	A3	A3	A3	-	-	-	-	9
talp_upc	-	A2	-	-	-	-	-	2
TMT	A1	A1	A1	-	-	A1	-	4
Transperfect	-	A3	-	-	-	A2	A2	7
Volctrans	-	-	-	-	-	-	A2	2
ZengHuiMT	-	-	-	-	-	-	A1	1
Total	10	9	10	5	2	6	16	58

Table 5: Overview of the submissions from all teams and test sets translating into English. We identify submissions to the abstracts test sets with an "A" and to the terminology test set with a "T". 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
Huawei_AGI	All	transformer (Own)	No	Yes	Yes (except zh2en)	No
Huawei_TSC	All	Marian,Fairseq	No	Yes	Yes (except en2de)	No
LISN	All	Fairseq	Yes	Yes	Yes	No
MT Learner	All	Marian-NMT	No	Yes	Yes	No
NVIDIA NeMo*	All	transformer (unspecified)	Yes	Yes	Yes	Yes
talp_upc	All	OpenNMT-transformer	Yes	No	Yes	No
nrpu-fjwu	All	Fairseq	Yes	Yes	No	No
TMT	All	Fairseq	Yes	No	Yes (except en2fr)	mBART (en2de,es,fr,ru)

Table 6: Overview of methods used by participating teams. Information is self-reported through the dedicated survey for each selected "best run" (information on the NVIDIA model is inferred from their system description (Subramanian et al., 2021)). BT indicates if backtranslation is used and LM if language models were used.

nificantly. Table 7 provide details of the in-domain data used by the teams.

For relevant language pairs, parallel data from other WMT tracks (e.g., 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.

We note that a number of the corpora used are referred to as "in house" corpus or data. This may indicate survey fatigue as this type of description is more frequently used for out of domain data, which appeared towards the end of the survey.

5 Automatic evaluation

For all the abstracts test sets, we evaluated system outputs using BLEU (Papineni et al., 2002) as provided by the Moses tool *mteval-v14.pl*[11]. We used this metric for the en2eu abstracts, summaries of animal experiments, and MEDLINE test sets. In en2zh, a modified version of the tool was used that removed white spaces and text is split in way that each character is a word.

The results for the en2eu abstract test sets are given in Table 9. There was a single team (Fujitsu DMATH) that submitted two runs, based on BPE dropout and sub-subword features with a Transformer (base) model. One of the runs (run2) included multilingual data from an English–Spanish terminology. The results are not as high as in the MEDLINE abstracts task, but they are above the baseline system, and they have improved from the best results from the 2020 challenge (0.1453 vs. 0.1279).

For the en2eu terminology test sets, we evaluated the translated concepts in terms of two metrics: (i) accuracy, by relying on strict matches (case insensitive) between the reference translation and predictions; and (ii) BLEU score, as measured by the Python NLTK module sentencebleu. The results are presented in Table 10. The same systems from FUJITSU DMATH participated in this task, and the BLEU score was higher than the score for abstracts, but there was a drop in performance from the results in 2020. This could have happened because the systems were tuned for abstracts and not terminologies. As is the case for abstracts, for the terminology set, run1 outperforms run2 again, showing that multilingual data seems to harm performance in this setting.

For the summaries of animal experiments, we only present the results obtained by our baseline system (Table 11).

Finally, for the Medline test sets, we performed evaluation based on all the sentences in the test set, including the poorly aligned ones, as well as an evaluation based on only the perfectly aligned ones (see Table 2). The results *from* English into the foreign languages are presented in Table 12, while the ones *into* English are presented in Table 13. The results calculated for all sentences, and not only the perfectly aligned ones, are published on the shared task's web site.[12]

For translation from English (cf. Table 12), the highest BLEU score of 0.5117 was obtained by the Transperfect team for en2es. Moreover, for all the language pairs for which the Huawei_TSC participated, i.e., en2de and en2zh, this team obtained the highest score, namely 0.3259 and 0.4650 respectively. For en2fr and en2it, the best performance was obtained by the Huawei_AGI team,

[11] https://github.com/moses-smt/mosesdecoder

[12] http://www.statmt.org/wmt21/results_biomedical.pdf

Language pair	team	Parallel corpus	size (sentence pairs)	Monolingual corpus	size (sentences)
de/en	Huawei_AGI	MEDLINE corpus supplied by WMT biomedical task organizers	2.4 M	Yes	53 M (en)
	Huawei_TSC	MEDLINE corpus supplied by WMT biomedical task organizers	3.03M	Yes	21.43M (en)
	nrpu-fjwu	sources provided by WMT biomedical task organizers (UFAL, Medline Abstracts and EMEA)	3.71 M	No	-
	TMT	corpus provided by WMT biomedical task organizers and UFAL.	2.5 M	Yes	2.5 M
es/en	Talp_upc	UFAL, Pubmed, Medline, IBECS,UNcor-m and OPUS	6.86 M	No	-
	TMT	corpus provided by WMT biomedical task organizers and UFAL.	1.6 M	Yes	1.6 M
	Transperfect	corpus provided by WMT biomedical task organizers	618 K	No	-
fr/en	Huawei_AGI	MEDLINE corpus supplied by WMT biomedical task organizers	3.6 M	Yes (en)	53 M
	LISN	Bio-medical corpora provided by the task organiser along with Taus and Cochrane	6 M	Yes (fr)	0.81 M
	nrpu-fjwu	sources provided by WMT biomedical task organizers e.g. UFAL, Scielo Health, EDP, Medline Titles, Medline Abstracts and EMEA.	4.36 M	No	-
	TMT	corpus provided by WMT biomedical task organizers and UFAL.	3.5 M	Yes	3.5 M
it/en	Huawei_AGI	MEDLINE corpus supplied by WMT biomedical task organizers, TAUS	374 K	Yes (en)	55 M
	MT Learner	Corpus supplied by WMT biomedical task organizers, and in-domain data filtered from an in-house corpus.	364 K	Yes (en)	1.5 M
pt/en	MT Learner	Corpus supplied by WMT biomedical task organizers, and in-domain data filtered from an in-house corpus.	1.6 M	Yes (en)	6.2 M
en/ru	MT Learner	Corpus supplied by WMT biomedical task organizers, augmented with in-house corpus.	2.2 M	Yes (en)	2.1 M
	NVIDIA	Corpus supplied by organizers, augmented with automatically filtered news-task corpus.	256k	?	?
	TMT	Corpus supplied by organizers, augmented with in-house corpus.	1 M	WMT biomedical task and UFAL	?
	Transperfect	"internal data" (unspecified)	6.1 M	No	-
en/zh	ECNU_PAHT	In-house corpus (unspecified)	6 M	No	-
	Huawei_AGI	In-house data collected from a portion of abstracts of China Master's and Doctoral Dissertations.	847 K	No	-
	Huawei_TSC	In-house corpus (unspecified)	1.35M	Yes	36.11M (zh), 21.43M (en)
	Transperfect	"internal data" (unspecified)	6.8 M	No	-

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).

with 0.4531 and 0.0.4425 respectively. Finally, the NVIDIA NeMo team obtained the best score (0.4139) for en2ru.

For translation into English (cf. Table 13), the highest score over all teams and language pairs was 0.5685, which was obtained by the MT Learner team for pt2en. TMT obtained the best results for three of the language pairs, namely de2en, es2en, and fr2en, with the scores 0.4501, 0.5382, and 0.4928 respectively. For it2en and zh2en, slightly higher scores (0.4570 and 0.3943 respectively) were obtained by the Huawei_AGI team, when compared to the ones from the MT learner (0.4558) and Huawei_TSC (0.3904) respectively. Finally, the NVIDIA NeMo team obtained the top score (0.4918) for the only language pair (ru2en) and run that they submitted.

6 Manual evaluation

Similar to previous years, we manually validated a sample of the abstracts to compare the teams' primary submissions to each other and to the reference translation.

For the MEDLINE abstracts, we aimed for approximately 100 perfectly aligned sentences and retrieved the corresponding abstracts. The sentences were randomly retrieved, but we aimed to select abstracts with a higher percentage of perfectly aligned

Language pair	team	Parallel corpus	size (sentence pairs)	Monolingual corpus	size (sentences)
en/de	Huawei_AGI	"in house data"	6 M	No	-
	Huawei_TSC	Corpus supplied by the WMT 2020 News task organizers	96.6 M	Yes	150M
	nrpu-fjwu	No	-	No	-
	TMT	Europarl-v10, Common Crawl corpus, ParaCrawl, News Commentary-v15 and Wiki Titles-v2	37.8 M	No	-
en/es	TALP	TED Talks	1.97 M	No	-
	TMT	Europarl-v7, Common Crawl corpus, News Commentary, ParaCrawl	30.3 M	No	-
	Transperfect	No	-	No	-
en/fr	Huawei_AGI	"in house data"	3 M	No	-
	LISN	WMT14 general domain corpus	35 M	No	-
	nrpu-fjwu	No	-	No	-
	TMT	Europarl-v7, Common Crawl corpus, News Commentary , English-French Giga Corpus	28 M	No	-
en/it	Huawei_AGI	"in house data"	6 M	No	-
	MT Learner	"in house data"	67 M	No	-
en/pt	MT Learner	"in house data"	67 M	No	-
en/ru	MT Learner	"in house data"	53 M	No	-
	NVIDIA	No?	-	No?	-
	TMT	Common Crawl corpus, News Commentary, ParaCrawl, Yandex Corpus, Wiki Titles-v2, Back-translated news	92 M	No	-
	Transperfect	No	-	No	-
en/zh	ECNU_PAHT	No	-	No	-
	Huawei_AGI	"in house data"	3 M	No	-
	Huawei_TSC	Corpus supplied by the WMT 2020 News task organizers	16.5M	Yes	150M
	Transperfect	No	-	No	-

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". (information on the NVIDIA model is inferred from their task paper).

Teams	Runs	BLEU
FJDMATH	run1	0.1453
	run2*	0.1403
Baseline	-	0.1091

Table 9: BLEU scores for the Abstract test set (en2eu). *Indicates the primary run as indicated by the participants.

Teams	Runs	Accuracy	BLEU
FJDMATH	run1	0.16	0.2783
	run2*	0.15	0.2674

Table 10: Scores for the Terminology test set (en2eu). *Indicates the primary run as indicated by the participants.

Teams	Runs	BLEU
Baseline	-	0.3800

Table 11: Performance scores for the test set of summaries of animal experiments (de2en).

sentences. This is the same strategy described in last year's publication (Bawden et al., 2020).

We only considered those teams which either submitted a publication to the workshop or filled in our survey with information about theirs runs. In some few cases, we could not considered some teams for the manual validation, e.g., MT learner for it2en, because the team filled in the survey when the manual validation was already been carried out.

For all teams, we considered the primary run (as indicated by the participants). The only exception was made for the Volctrans team, for which we considered the run with the highest BLEU score, according to the automatic evaluation. The primary runs that we considered in the manual validation are listed below:

- en2de (3 teams): Huawei_AGI (run3), Huawei_TSC (run3), TMT (run1)
- en2es (3 teams): talp_upc (run2), TMT (run1), Transperfect (run2)
- en2fr (3 teams): Huawei_AGI (run3), LISN (run1), TMT (run1)

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh
ECNU_PAHT	run1	-	-	-	-	-	-	0.4197
	run2	-	-	-	-	-	-	0.4364
	run3	-	-	-	-	-	-	0.4504*
Haozhiweizi	run1	-	-	-	-	-	-	0.4381*
Huawei_AGI	run1	0.3172	-	0.4531	0.4301	-	-	0.4342
	run2	0.3198	-	0.4424	0.4334	-	-	0.4440
	run3	0.3172*	-	0.4489*	0.4425*	-	-	0.4293*
Huawei_TSC	run1	0.3259	-	-	-	-	-	0.4639
	run2	0.3329	-	-	-	-	-	0.4640*
	run3	0.3259*	-	-	-	-	-	0.4650
JingDong	run1	-	-	-	-	-	-	0.3970*
LISN	run1	-	-	0.3912*	-	-	-	-
	run2	-	-	0.3913	-	-	-	-
	run3	-	-	0.4293	-	-	-	-
NVIDIA NeMo	run1	-	-	-	-	-	0.4139	-
	run2	-	-	-	-	-	0.4112*	-
talp_upc	run1	-	0.4084	-	-	-	-	-
	run2	-	0.4142*	-	-	-	-	-
TMT	run1	0.2765	0.4354	0.4456	-	-	0.3289	-
Transperfect	run1	-	0.5117	-	-	-	0.3686	0.4029
	run2	-	0.5012*	-	-	-	0.3492*	0.4025*
	run3	-	0.4917	-	-	-	-	-
Volctrans	run1	-	-	-	-	-	-	0.4406
	run2	-	-	-	-	-	-	0.4433
	run3	-	-	-	-	-	-	0.4361*
ZengHuiMT	run1	-	-	-	-	-	-	0.4126
Baseline	-	0.2536	0.4027	0.3924	0.4147	0.4304	0.2451	0.3096

Table 12: BLEU scores for "OK" aligned test sentences, from English. For the Volctrans team, we renamed the runs: run1=run1, run2=nnmt, run3=nnmtne. *Indicates the primary run as indicated by the participants.

- en2it (1 team): Huawei_AGI (run3)
- en2ru (3 teams): NVIDIA NeMo (run2), TMT (run1), Transperfect (run2)
- en2zh (6 teams): ECNU_PAHT (run3), Haozhiweizi (run1), Huawei_AGI (run3), Huawei_TSC (run2), Transperfect (run2), Volctrans (run2)
- de2en (4 teams): Huawei_AGI (run3), Huawei_TSC (run3), nrpu-fjwu (run1), TMT (run1)
- es2en (4 teams): nrpu-fjwu (run1), talp_upc (run2), TMT (run1), Transperfect (run2)
- fr2en (4 teams): Huawei_AGI (run3), LISN (run3), nrpu-fjwu (run1), TMT (run1)
- it2en (2 teams): Huawei_AGI (run3)
- pt2en (1 team): MT Learner (run1)
- ru2en (4 teams): NVIDIA NeMo (run1), TMT (run1), Transperfect (run2)
- zh2en (6 teams): ECNU_PAHT (run3), Haozhiweizi (run1), Huawei_AGI (run3), Huawei_TSC (run3), Transperfect (run2), Volctrans (run2)

For each language pair, we generated pairwise combinations of either two teams' primary runs or one primary run and the reference translation. The evaluator first compared pairs of sentences, followed by whole abstracts; the exception was en2zh and zh2en, where only whole abstracts were compared due to the otherwise infeasible large amount of evaluation required. These pairs of translations were manually validated in the Appraise tool (Federmann, 2010) following the same procedure carried out in previous years. For each pair of sentence or abstracts, the aim of the evaluation was to decide whether the translations were of equivalent quality or whether one was better than the other. The results of the manual validation are presented in

Teams	Runs	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en
ECNU_PAHT	run1	-	-	-	-	-	-	0.3232
	run2	-	-	-	-	-	-	0.3232
	run3	-	-	-	-	-	-	0.3546*
Haozhiweizi	run1	-	-	-	-	-	-	0.3713*
Huawei_AGI	run1	0.3956	-	0.4860	0.4570	-	-	0.3943
	run2	0.4132	-	0.4871	0.4569	-	-	0.3785
	run3	0.4048*	-	0.4871*	0.4550*	-	-	0.3934*
Huawei_TSC	run1	0.4230	-	-	-	-	-	0.3828
	run2	0.4258	-	-	-	-	-	0.3921
	run3	0.4310*	-	-	-	-	-	0.3904*
JingDong	run1	-	-	-	-	-	-	0.3041*
LISN	run1	-	-	0.4322	-	-	-	-
	run2	-	-	0.4112	-	-	-	-
	run3	-	-	0.4325*	-	-	-	-
MT Learner	run1	-	-	-	0.4558*	0.5584*	0.4871*	-
	run2	-	-	-	0.4548	0.5685	0.4751	-
NVIDIA NeMo	run1	-	-	-	-	-	0.4918	-
nrpu-fjwu	run1	0.3524*	0.4590*	0.3840*	-	-	-	-
	run2	0.3495	0.4598	0.3921	-	-	-	-
	run3	0.3367	0.4600	0.3772	-	-	-	-
talp_upc	run1	-	0.4194	-	-	-	-	-
	run2	-	0.4194*	-	-	-	-	-
TMT	run1	0.4501	0.5382	0.4928	-	-	0.4061	-
Transperfect	run1	-	0.5237	-	-	-	0.4794	0.3291
	run2	-	0.4991*	-	-	-	0.4769*	0.3212*
	run3	-	0.4969	-	-	-	-	-
Volctrans	run1	-	-	-	-	-	-	0.2911
	run2	-	-	-	-	-	-	0.3796
ZengHuiMT	run1	-	-	-	-	-	-	0.2832
Baseline	-	0.3392	0.3959	0.3796	0.4075	0.4506	0.3115	0.2237

Table 13: BLEU scores for "OK" aligned test sentences, into English. For the Volctrans team, we renamed the runs: run1=base, run2=nnmt. *Indicates the primary run as indicated by the participants.

various tables as summarized below:

- pt2en: Table 14
- en2es and es2en: Table 15
- en2de and de2en: Table 16
- en2fr and fr2en: Table 17
- en2it and it2en: Table 18
- en2zh and zh2en: Table 19
- en2ru and ru2en: Table 20

We identified the item (a system or the reference translation) of each pairwise comparison that performed better (see respective tables) and ran a Wilcoxon Signed-Rank Test from the Python `scipy` library. We consider all comparisons for two particular items over all validated abstracts and sentences, except for skipped ones. The test was calculated for the abstracts and the sentences. We mark in bold in the respective tables the ones that were found to be significant (i.e., p-value< 0.05) and otherwise the systems are considered to be similar. We considered one item superior than the other when either the validation of the abstract of the sentences was statistically significant. For the language pairs validated by two experts (i.e., es2en and pt2en), we only considered one item to be superior than the other when at least two of the four comparisons (2x for the abstracts, 2x for the sentences) were statistically significant.

We ranked the system by assigning points to each item: 3 points if superior to the opponent, 1 point when they have similar quality, and no points if inferior to the opponent. Based on the sum of these points over all comparisons, we ranked the systems and the reference translations as shown be-

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
pt2en	reference-MT Learner	14	3	9	2	112	6	92	14

Table 14: Manual validation for the pt2en MEDLINE abstracts test set. The test set could only be validated with regard to the content of the translation, but not regarding the quality of the English translations.

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
en2es	TMT-**reference**	8	0	0	**8**	103	9	17	**77**
	TMT-**talp_upc**	8	0	0	**8**	103	6	23	**74**
	TMT-**Transperfect**	8	0	0	**8**	103	3	21	**79**
	reference-talp_upc	8	3	4	1	103	17	77	9
	reference-Transperfect	8	1	7	0	103	8	90	5
	talp_upc-**Transperfect**	8	1	3	4	103	8	75	**20**
es2en	reference-nrpu-fjwu	13/4	7/2	3/1	3/1	107/31	23/**14**	53/13	31/4
	reference-**TMT**	13/4	0/2	4/1	**9**/1	107/31	6/13	57/10	**44**/8
	reference-**Transperfect**	13/4	1/3	4/0	**8**/1	107/31	10/**15**	60/11	**37**/5
	reference-talp_upc	13/4	**9**/3	2/0	2/1	107/31	33/12	49/14	25/5
	nrpu-fjwu-**TMT**	13/4	1/0	2/2	**10**/2	107/31	5/3	67/24	**35**/4
	nrpu-fjwu-**Transperfect**	13/4	0/1	3/1	**10**/2	107/31	4/6	67/22	**36**/3
	nrpu-fjwu-talp_upc	13/4	9/2	1/2	3/0	107/31	31/9	53/17	23/5
	TMT-Transperfect	13/4	3/2	9/2	1/0	107/31	14/**8**	84/22	9/1
	TMT-talp_upc	13/4	**10**/3	3/0	0/1	107/31	**42**/8	61/17	4/6
	Transperfect-talp_upc	13/4	**10**/3	2/0	1/1	107/31	**36**/6	63/19	8/6

Table 15: Manual validation for the en2es and es2en MEDLINE abstracts test set. The better performing MT system (or reference translation) in each pairwise comparison is shown in bold, as well as the respective value that has been identified as superior. For the es2en test set, the values on the left are the validation with regard to the content of the translations, while the ones on the right are regarding the quality of the English translations.

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
en2de	reference-TMT	9	5	2	1	114	40	38	35
	reference-Huawei_AGI	9	**6**	2	0	114	**51**	40	21
	reference-Huawei_TSC	9	**4**	4	0	114	13	57	**43**
	TMT-Huawei_AGI	9	2	4	2	114	**36**	60	16
	TMT-**Huawei_TSC**	9	0	4	**4**	114	3	59	**51**
	Huawei_AGI-**Huawei_TSC**	9	0	1	**7**	114	5	33	**74**
de2en	**Huawei_TSC**-reference	11	**9**	1	1	93	**31**	44	17
	Huawei_TSC-Huawei_AGI	11	**9**	1	1	93	**38**	50	5
	Huawei_TSC-nrpu-fjwu	11	**11**	0	0	93	**58**	33	2
	Huawei_TSC-TMT	11	6	2	3	93	14	69	10
	reference-Huawei_AGI	11	7	1	3	93	**40**	30	23
	reference-nrpu-fjwu	11	**9**	1	1	93	**47**	28	18
	reference-TMT	11	5	1	5	93	22	47	24
	Huawei_AGI-nrpu-fjwu	11	**10**	1	0	93	**44**	35	14
	Huawei_AGI-**TMT**	11	1	6	4	93	9	56	**28**
	nrpu-fjwu-**TMT**	11	0	2	**9**	93	5	43	**45**

Table 16: Manual validation for the en2de and de2en MEDLINE abstracts test set. The better performing system (or reference translation) in each pairwise comparison is shown in bold, as well as the respective value that has been identified as superior.

low (the points obtained are shown in parentheses):

- en2de: Huawei_AGI (0) < TMT (4) = reference (5) < Huawei_TSC (7)
- en2es: TMT (0) < reference (4) = talp_upc (4) < Transperfect (7)
- en2fr: Huawei_AGI (2) = TMT (2) < LISN (5) < reference (6)
- en2it: Huawei_AGI (1) = reference (1)

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
en2fr	**reference**-LISN	16	10	1	3	100	**58**	13	18
	reference-Huawei_AGI	16	**14**	0	2	100	**65**	18	17
	reference-TMT	16	**14**	1	1	100	**65**	18	17
	LISN-Huawei_AGI	16	6	1	7	100	37	29	23
	LISN-TMT	16	4	4	6	100	30	30	29
	Huawei_AGI-TMT	16	7	7	2	100	29	43	28
fr2en	nrpu-fjwu-**Huawei_AGI**	12	2	0	**10**	79	15	19	**42**
	nrpu-fjwu-LISN	12	4	1	7	79	19	26	33
	nrpu-fjwu-reference	12	3	1	8	79	28	13	37
	nrpu-fjwu-**TMT**	12	1	0	**11**	79	9	24	**45**
	Huawei_AGI-LISN	12	7	0	5	79	27	30	21
	Huawei_AGI-reference	12	7	1	4	79	**37**	20	21
	Huawei_AGI-TMT	12	3	3	6	79	17	36	25
	LISN-reference	12	6	2	4	79	31	26	21
	LISN-TMT	12	3	0	9	79	16	36	26
	reference-**TMT**	12	3	2	7	79	19	18	**41**

Table 17: Manual validation for the en2fr and fr2en MEDLINE abstracts test set. The better performing system (or reference translation) in each pairwise comparison is shown in bold, as well as the respective value that has been identified as superior.

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
en2it	Huawei_AGI-reference	10	6	0	4	100	38	35	27
it2en	Huawei_AGI-reference	11	4	0	7	102	32	40	30

Table 18: Manual validation for the en2it and it2en MEDLINE abstracts test sets. For the it2en test set, only the translation from Italian into English was assessed, but not the quality of the English text.

- en2ru: NVIDIA Nemo (3) = reference (3) = TMT (3) = Transperfect (3)
- en2zh: ECNU_PAHT (3) < Huawei_AGI (4) = Haozhiweizi (4) = Transperfect (4) < Huawei_TSC (9) < Volctrans (12) < reference (16)
- de2en: nrpu-fjwu (0) < Huawei_AGI (3) < reference (7) < TMT (8) < Huawei_TSC (10)
- es2en: nrpu-fjwu (2) = reference (2) = talp_upc (2) < TMT (10) = Transperfect (10)
- fr2en: nrpu-fjwu (2) = reference (2) < LISN (4) < Huawei_AGI (8) = TMT (8)
- it2en: reference (1) = Huawei_AGI (1)
- pt2en: reference (1) = MT Learner (1)
- ru2en: TMT (0) < Transperfect (4) < reference (5) < NVIDIA NeMo (7)
- zh2en: Huawei_AGI (5) < Haozhiweizi (6) = Volctrans (6) = Huawei_TSC (6) = ECNU_PAHT (6) = reference (6) < Transperfect (8)

Abstracts for en2eu (Osagaiz + Gaceta) were manually validated following the same approach, but only at the sentence level. As there was one submission for this language pair, we only generated a pairwise combination of the participant's run and the reference. The run with the highest BLEU score was selected for validation:

- en2eu (1 team): FJDMATH (run1)

The translations were evaluated by three annotators using the Appraise tool, and the averaged results are presented in Table 21. The ranking based on the points is as follows:

- en2eu: FJDMATH (0) < reference (3)

7 Discussion

7.1 Quality of the MT evaluation process.

Marie et al. (2021) introduced guidelines for the evaluation of MT quality, based on four criteria:

Language	Pair	Abstracts			
		Total	**A>B**	**A=B**	**A<B**
en2zh	ECNU_PAHT-Huawei_AGI	13	3	1	9
	ECNU_PAHT-Transperfect	13	5	0	8
	ECNU_PAHT-**Volctrans**	13	1	0	**12**
	ECNU_PAHT-Haozhiweizi	13	4	3	6
	ECNU_PAHT-**Huawei_TSC**	13	1	2	**10**
	ECNU_PAHT-**reference**	13	0	1	**12**
	Huawei_AGI-Transperfect	13	8	1	4
	Huawei_AGI-Volctrans	13	2	3	8
	Huawei_AGI-Haozhiweizi	13	7	1	5
	Huawei_AGI-**Huawei_TSC**	13	2	1	**10**
	Huawei_AGI-**reference**	13	2	2	**9**
	Transperfect-**Volctrans**	13	2	1	**10**
	Transperfect-Haozhiweizi	13	7	1	5
	Transperfect-Huawei_TSC	13	3	0	10
	Transperfect-**reference**	13	2	1	**10**
	Volctrans-Haozhiweizi	13	**10**	1	2
	Volctrans-Huawei_TSC	13	3	6	4
	Volctrans-reference	13	3	2	8
	Haozhiweizi-Huawei_TSC	13	4	1	8
	Haozhiweizi-**reference**	13	2	0	**11**
	Huawei_TSC-**reference**	13	2	1	**10**
zh2en	Haozhiweizi-Volctrans	19	11	1	7
	Haozhiweizi-Huawei_TSC	19	10	3	6
	Haozhiweizi-reference	19	9	4	5
	Haozhiweizi-ECNU_PAHT	19	10	2	7
	Haozhiweizi-Huawei_AGI	19	10	5	4
	Haozhiweizi-Transperfect	19	4	6	9
	Volctrans-Huawei_TSC	19	3	8	8
	Volctrans-reference	19	7	3	8
	Volctrans-ECNU_PAHT	19	8	4	7
	Volctrans-Huawei_AGI	19	9	3	7
	Volctrans-Transperfect	19	5	3	11
	Huawei_TSC-reference	19	6	5	7
	Huawei_TSC-ECNU_PAHT	19	9	2	8
	Huawei_TSC-Huawei_AGI	19	10	1	8
	Huawei_TSC-Transperfect	19	7	4	8
	reference-ECNU_PAHT	19	12	0	6
	reference-Huawei_AGI	19	9	2	7
	reference-Transperfect	19	7	4	7
	ECNU_PAHT-Huawei_AGI	19	8	4	7
	ECNU_PAHT-Transperfect	19	3	6	10
	Huawei_AGI-**Transperfect**	19	3	3	**13**

Table 19: Manual validation for the en2zh and zh2en MEDLINE abstracts test set. Only the abstracts were validated. The better performing system (or reference translation) in each pairwise comparison is shown in bold, as well as the respective value that has been identified as superior.

(1) use of an evaluation method in addition to/in lieu of BLEU, (2) use of statistical significance testing to compare systems, (3) direct computation of scores instead of copying from previous experiments and (4) comparison of systems only if the same training, validation and test sets have been used, as well as the same pre-processing steps.

The evaluation carried out in this task is compliant with criteria (1-3). However, participants are free to use their choice of training corpus, validation corpus and pre-processing methods. This approach was selected to foster participant creativity and set a lower entry cost to the task. It is a limitation in the comparability of the systems submitted for this task. As a mitigation strategy, we encourage participants to also submit detailed descriptions of system particulars to provide transparency on the material and methods used.

A future edition of the task could introduce a "constrained" track where pre-processed training/validation sets would be supplied to be used exclusively (as in the WMT news translation task (Barrault et al., 2020)).

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
en2ru	TMT-Transperfect	16	6	1	9	95	15	61	19
	TMT-NVIDIA NeMo	16	4	1	11	95	20	47	25
	TMT-reference	16	6	0	10	95	27	42	25
	Transperfect-NVIDIA NeMo	16	6	5	5	95	26	52	15
	Transperfect-reference	16	4	6	6	95	22	52	21
	NVIDIA NeMo-reference	16	2	9	5	95	13	64	15
ru2en	**Transperfect**-TMT	16	**10**	6	0	109	**42**	56	9
	Transperfect-reference	16	2	9	5	109	25	65	19
	Transperfect-**NVIDIA NeMo**	16	0	10	**6**	109	7	87	15
	TMT-**reference**	16	0	3	**13**	109	10	56	**41**
	TMT-**NVIDA NeMo**	16	0	1	**15**	109	4	61	**42**
	reference-NVIDIA NeMo	16	2	9	5	109	16	71	22

Table 20: Manual validation for the en2ru and ru2en MEDLINE abstracts test sets. The better performing system (or reference translation) in each pairwise comparison is shown in bold, as well as the respective value that has been identified as superior.

Language	Pair	Sentences			
		Total	**A>B**	**A=B**	**A<B**
en2eu	reference-FJDMATH	100	**61**	16	23

Table 21: Manual validation for the en2eu (Osagaiz and Gaceta) abstracts test set. The better performing system (or reference translation) in each pairwise comparison is shown in bold, as well as the respective value that has been identified as superior. The values show the validation performed by the Basque native speakers (averaged over three annotators).

7.2 Quality of the system translations

We report below some of the major observations collected throughout the manual validation of the selected runs and the reference translations.

7.2.1 MEDLINE test sets

de (from en) The perceived quality of translations was high, with a high proportion of perfect or near perfect translations. The translation quality between participating systems differ only by small nuances. For example, translations differ only by word order or use different synonyms for a specific medical term. All participating systems had problems translating abbreviations. For instance, "image quality (IQ)" becomes "Bildqualität (IQ)" instead of BQ. One participant could not generate umlauts (e.g., ä,ö,ß, ...) and another participant produced only lowercased text. Both problems lead to slightly reduced quality of translations.

en (from de) Overall, the translation quality was high. Most translated sentences were understandable, except in cases where the original German sentences were too long to translate correctly. In some cases, the translations captured the intended meaning, but took information from different sentences, or even used synonyms, which were not direct literal translations of words, such as "neurosensory retina" for "macular region". There were also some translations from the first person point of view, rather than the impersonal, for a more personal touch. If such examples represent MT outputs rather than the reference translations, the quality of translation is approaching native speaker level.

Some texts contained small errors, which should be easy to avoid, such as capitalization of the first word after "e.g.", the use of lower case letters for well-known abbreviations like "AR" for augmented reality, proper nouns ("Marburg heart score"), and gene names (PD-1). Also a repetition of words in common expressions like the German *wie z. B.* should not have been translated as "e.g. For example". Using the same word twice in a sentence could have been avoided: "Relapse is defined as the recurrence" instead of "Recurrence is defined as the recurrence". Interestingly, a translation actually corrected a capitalization error in the original German text, from *l. reuteri* to *L. reuteri*, for the genus *Lactobacillus* in the scientific name of the bacteria.

The correct translation of medical terms also

proved to be difficult in cases such as "centrum" instead of "ventrum", "neurology" instead of "urology", or "endourolgy" instead of "endourology". Medical terms pertaining to a specific field, were also difficult to translate properly, such as the German *Pluszeichen* to the English "plus disease" in the context of retinal disease and "fusion biopsy" to describe the method of using both magnetic resonance imaging and ultrasound to take prostate biopsies.

es (from en) Quality of translations is improving every year and in many cases it is difficult to identify which translations are machine made.

There have been cases in which abbreviations were not translated correctly (e.g. *HGS* vs *FPM* for *fuerza de prensión manual*) and some times specific terms were not translated, e.g. *receiver operating characteristic curve*. Only in a few cases word gender was different to the article one.

There are examples of words that are not translated properly for instance *adnexal* has been translated as *adnexiales* instead of *anexas* by one of the teams.

In addition to individual sentences, the manual evaluation included abstracts as well. Since translations were sentence based, there were cases in which there were misaligned information between the content of the sentences in the abstract, even if the translated sentences were perfectly fine in their own.

We identified that a team might have been missing accents on vowels and special letters such as *ñ*, which seemed to indicate that the translation was machine made.

fr (from en) The overall quality of translation was high. We noted that many of the sentences compared were identical or nearly identical. In many cases, the translations selected as superior were chosen based on small nuances, such as capitalization, typography, ordering of words or sentence structure that appeared more adequate to a native speaker, while causing no difference in the understanding of the text. A few terms or acronyms were sometimes untranslated but the resulting text could be understood (for example, *incidentaloma* was used instead of *incidentalome*). Erroneous disambiguation was observed in some translations for one abstract discussing pressure at the fingertips, where *numérique* was used instead of *digital*. At the abstract level, some vocabulary consistency issues could be evidenced. For example, one translation used the synonyms *sclérodermie systémique* and *sclérose systémique* alternatively as translation for "systemic sclerosis". While individual sentences were correctly translated, it created confusion at the abstract level, compared to the "reference" translation, which used the term *sclérodermie* throughout. Consistently with the 2020 edition, arbitration between sentences exhibiting a fluency or grammatical flaw vs. a semantic or clinical flaw was conducted as favoring the semantic or clinical correctness. However, the nature of the "reference" translation (which is often not produced by professional translators and does not necessarily provide straight forward sentence-by-sentence translations (Névéol et al., 2020)) introduces bias and difficulty in the evaluation: highly fluent text with some semantic distance with the "original" sentence to be translated can sometimes be easily identified as the reference text. It is difficult to arbitrate between this high quality text and the machine translation that will attempt to be semantically closer while exhibiting language flaws.

en (from fr) Translation quality was generally very high, with some variation in the quality depending on the topic of the abstract being translated (most systems struggled with the more literary text from a sociology abstract). This meant that many of the decisions were, as with the other language pairs, based on preferences and formatting rather than differences in meaning (punctuation, capitalisation, minor grammar mistakes).

The most serious errors observed were with the translation of specific terms, such as illnesses and drugs. They were particularly prevalent for acronyms, which were sometimes not translated and sometimes poorly translated (another more common acronym being used instead, e.g. *ADHD* instead of *ADPKD*). A few tricky sentences revealed the risk of major semantic errors resulting from seemingly small and localised errors. We give two such examples here. Firstly, concerning temporality, several systems translated French *puis* 'then' as English *and* in *un anticoagulant puis l'aspirine* 'an anti-coagulant then aspirine', a sentence for which the order in which drugs are given may be fundamental. Secondly, many systems stumbled on the translation of French *cela ne s'accompagne pas d'une attention égale au rôle de l'écoute* 'this is not accompanied by equal attention to the role of listening', inverting the order of the two underlined

words, resulted in the opposite meaning (i.e. listening receiving more rather than less attention).

As noted above for en2fr, despite very high MT quality, reference translations are often still easily identifiable due to them being less literal. This means that they are often characterized by better word choice and more natural syntax, but it can also mean that they are less adequate because of missing information or even additional details not present in the French text.

it (from en) The quality of the translation was on average high, probably higher than 2020. Some of the sentences compared were almost identical.

From a terminological viewpoint, it is possible to identify some inaccuracies in the choice of translating terms in the target language. For example, the term 'malignancies' was translated by one system with *tumori* (corresponding to the English 'tumors') having a broader meaning than *neoplasie maligne* ('malignant neoplasms'). Furthermore, cases of erroneous choices of the translating terms can be identified. The adjective 'unpreventable' was translated as *non prevedibile* ('unpredictable') instead of *non evitabile*, thus causing the transmission of an incorrect information in the target text.

Another error can be identified in the choice of the generic verb 'consider'. In the case of the sentence 'total laryngectomy should be considered [...]', the construction was wrongly translated as *la laringectomia dovrebbe essere considerata il trattamento di scelta*, that is the 'Laryngectomy should be considered the treatment of choice'. It is also possible to identify cases of non-translation in the Italian text (for example the name of the city 'Zurich' remained untranslated) and the presence of anglicisms as *imaging diagnostico* chosen as the translation of 'diagnostic imaging', although the Italian equivalent *diagnostica per immagini* is commonly used in the target language.

Moreover, the term 'livestock' was translated with *mandria* 'herd', *bestiame* 'livestock' or *allevamento* 'farm'. One interesting case was the term 'blacks', which was translated with *non bianchi* (non-whites) instead of the more frequently used *neri*.

Finally, from a syntactic point of view, there were a couple of examples where the syntactic tree was built erroneously: for example, the phrase 'incidental thyroid cancer rates' was translated as *i tassi di carcinoma tiroideo incidentale* ('rates of incidental thyroid cancer'); another example is '4 cm lobule contoured mass' translated as *massa sagomata di 4 cm del lobulo* ('4 cm contoured mass of the lobule').

zh (from en) The quality of translation was high overall. The primary reason for awkward translations was word order, since English and Chinese employ different word orders not only at the word level, but also at the phrase level. Consider this example, where the source text was *surveillance and early warning of infectious diseases in China*. A good translation first needed to adjust word order within *infectious diseases in China* to yield 中国传染病 (the order is *China* then *infectious diseases*). Then phrase order also needed to be adjusted to yield 中国传染病监测预警能力 (the order is *China infectious diseases* then *surveillance and early warning*). Some translations failed to make these necessary adjustments, such that the Chinese translation in the original English word order rendered the translation awkward or even unintelligible.

Another source of deficiency was translations that were too literal. For instance, *under-reporting* was most often translated as 报告不足 (*insufficient reporting*), though a more native, conventional wording would actually be 漏报 (*omitted reporting*). Consider another example, *appraised persons* in the context of a study of familial relationships. While the reference translation 被鉴定人 was the most fitting, some teams' translation 被评估者 (a person to be evaluated) was also a good fit. However, another translation such as 评估人员 (evaluation personnel) was outright incorrect, as the meaning went from "a passive person being evaluated" to "an active person evaluating someone else".

en (from zh) The quality of translation was also high and noticeably better than last year. Where some translations last year were unintelligible, such cases have disappeared this year. In addition, there was a range of translation qualities last year, but this year every team's translation quality was high.

This year, the aspects that distinguish a better translation from a worse one are more subtle. Firstly, Chinese sentences as delimited by the Chinese full stop "。" are often equivalent to short paragraphs in English. A good translation should therefore split a Chinese sentence where necessary into multiple English sentences. Secondly, a technical term may have synonyms (e.g. *acetabulum*

labrum and *acetabular lip*), and a better translation should use one synonym consistently within the same abstract instead of mixing different ones. Thirdly, a good translation should use the most fitting wording, a task that requires good understanding of sentence context as well as domain knowledge. Consider this example, where 夫妻婚姻关系 (*marital relationship of married couples*) and 双向关联性 (*bi-directional correlation*) occur within the same sentence. A good translation can tease out *relationship* and *correlation*, but a worse translation simply uses *relationship* for both occurrences. This year, the better translations achieved the above three aspects, though rarely all three at once in the same translation.

en (from pt) While there was no significant difference between the automatic translations from the MT Learner team and the reference translation, we highlight some situations in which one was considered better than the other. On one hand, some mistakes were very subtle, such as a typo in a word (e.g., "verage" instead of "average"), or an inappropriate capitalization of a word. On the other hand, there were some semantic mistakes related to the translation of the sentences. For instance, the passage "insuficiência do glúteo médio esteve presente em todos os sujeitos" was translated as "the gluteus medius was insufficient in all patients".

Overall Based on these comments, many language pairs would benefit from a visual feature highlighting differences in the translations in the interface to focus the analysts' attention on often small differences. It could also be relevant to focus manual evaluation on a number of targeted linguistic features that seem to remain difficult (based on 2020 and 2021 observations), such as (a) translation of acronyms (b) vocabulary/grammatical consistency throughout a document (c) translation of numerical data. This might help make the manual evaluation more comparable between language pairs. However, it raises the question of the method to use for the selection of sentences/passages exhibiting the desired phenomena.

7.2.2 Osagaiz/Gaceta abstract test sets (en2eu)

In general, despite the fact that in the manual evaluation FJDMATH ranked below the references, the translations generated by the system were good, containing sentences with high-level of fluency and high adequacy with respect to the source. Similar to what has been observed in other language pairs, the system sometimes struggled with the translation of acronyms. For example, "non-motor symptoms (NMS)" should be translated to "sintoma ez motor (SEM)" but the participant's system translated it as "sintoma ez-motor (NMS)"; or "amiotrophic lateral sclerosis (ALS)" should be "alboko esklerosi amiotrofikoa (AEA)".

On the other hand, sometimes the reference translation in Basque contained extra information that was not present in the source English sentence. In these cases the additional information is contained in the context (i.e. surrounding sentences in the abstract). This can penalize the BLEU score of a correct sentence-level translation. For example, the source sentence "*Important hormonal changes happen during pregnancy and lactation*" was correctly translated by the system to "*Hormona aldaketa garrantzitsuak gertatzen dira haurdunaldian eta laktazioan.*". However, the reference translation also mentions "physiological changes in the body" (i.e. "*Haurdunaldi eta edoskitzaroan zehar,* ***gorputzeko maila askotan aldaketa fisiologikoak eragingo dituzten*** *gorabehera hormonal nabariak gertatuko dira*"). Document or abstract level translation systems could potentially alleviate this problem by leveraging contextual information from surrounding sentences.

7.2.3 Terminology test sets (en2eu)

The participating team (FJDMATH) had more difficulty in the translation of ICD-10 code descriptions (16% accuracy), particularly if we compare them with the results obtained by many teams last year (∼70% accuracy). Note that ICD-10 codes included in the test sets every year are different, but the performance difference is big considering there was more in-domain training data available this year. Some of the common mistakes observed in the system are word repetition (e.g. *hortz-posizioko anomaliak, hortz edo hortz guztiz eruptatuen posizioa* - "tooth position anomalies, tooth and tooth"), not translating an English word (e.g. *tidal-wave* instead of *olatu erraldoi*) or low adequacy (i.e. *huts egin du jaioberrian irabaztean* / (en) *missed when winning the newborn* where it should be *jaioberriaren garapeneko atzerapen* / (en) *Failure to thrive in newborn*). It is possible that the system was not sufficiently fine-tuned for the technical and specific language employed in ICD-10 code descriptions.

8 Conclusions

Our sixth edition of the WMT Biomedical Translation addressed a total of eight language pairs and three types of documents. One more time, we could assess the performance of current MT technology for the translation of biomedical textual resources. Further, we could attract the attention of many teams and received submissions for most of our test sets.

Similar as in the more recent editions of the shared task, participating system could perform better than the reference translation for many of the language pairs. However, this is still a challenge for en2zh. In future editions of this challenge, we aim at releasing more resources, especially additional training data, adding new language pairs, and considering a variety of test sets.

Acknowledgements

We would like the thank the participants Virginia Adams (NVIDIA NeMo), Cao Jun (Volctrans/ByteDance), Wei Peng (Huawei_AGI), and multiple individuals in the Huawei_TSC team for supporting us in the manual validation. The Academy of Medical Sciences of Bilbao and the Gaceta Médica de Bilbao have collaborated in WMT 2021 by providing us with their documents.

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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.

Bettina Bert, Antje Dörendahl, Nora Leich, Julia Vietze, Matthias Steinfath, Justyna Chmielewska, Andreas Hensel, Barbara Grune, and Gilbert Schönfelder. 2017. Rethinking 3R strategies: Digging deeper into AnimalTestInfo promotes transparency in in vivo biomedical research. *PLOS Biology*, 15(12):1–20.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, 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.

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.

DA Lindberg, BL Humphreys, and AT McCray. 1993. The unified medical language system. *Yearb Med Inform*, 1:41–51.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Ander Martínez. 2021. The Fujitsu DMATH submissions for WMT21 News Translation and Biomedical Translation Tasks. In *Proceedings of the Sixth Conference on Machine Translation: Shared Task Papers*.

Sumbal Naz, Sadaf Abdul Rauf, and Sami Ul Haq. 2021. FJWU participation for the WMT21 Biomedical Translation Task. In *Proceedings of the Sixth Conference on Machine Translation: Shared Task Papers*.

Aurélie Névéol, Antonio Jimeno Yepes, and Mariana Neves. 2020. MEDLINE as a parallel corpus: a survey to gain insight on French-, Spanish- and Portuguese-speaking authors' abstract writing practice. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3676–3682, Marseille, France. European Language Resources Association.

Mariana Neves, Daniel Butzke, Antje Dörendahl, Nora Leich, Benedikt Hummel, Gilbert Schönfelder, and Barbara Grune. 2019. Overview of the CLEF eHealth 2019 Multilingual Information Extraction. In *Tenth International Conference of the CLEF Association (CLEF 2019)*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Bardia Rafieian and Marta Ruiz Costa-Jussà. 2021. High frequent in-domain words segmentation and forward translation for the WMT21 Biomedical task. In *Proceedings of the Sixth Conference on Machine Translation: Shared Task Papers*.

Sandeep Subramanian, Oleksii Hrinchuk, Virginia Adams, and Oleksii Kuchaiev. 2021. NVIDIA NeMo's Neural Machine Translation Systems for English ↔ German and English ↔ Russian News and Biomedical Tasks at WMT21. In *Proceedings of the Sixth Conference on Machine Translation: Shared Task Papers*.

Weixuan Wang, Wei Peng, Xupeng Meng, and Qun Liu. 2021a. Huawei AARC's submissions to the WMT21 Biomedical Translation Task: Domain Adaption from a Practical Perspective. In *Proceedings of the Sixth Conference on Machine Translation: Shared Task Papers*.

Xing Wang, Zhaopeng Tu, and Shuming Shi. 2021b. Tencent AI Lab Machine Translation Systems for the WMT21 Biomedical Translation Task. In *Proceedings of the Sixth Conference on Machine Translation: Shared Task Papers*.

Sarah Wilde. 2021. African languages to get more bespoke scientific terms. *Nature (news)*, pages 469–470.

Hao Yang, ZhanglinWu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Chuanfei Xu, Min Zhang, and Ying Qin. 2021. HW-TSC's submissions to the WMT21 Biomedical Translation Task. In *Proceedings of the Sixth Conference on Machine Translation: Shared Task Papers*.