Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages

Rachel Bawden, Giorgio Di Nunzio, Cristian Grozea, Iñigo Unanue, Antonio Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, et al.

Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages

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

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

Abstract

Machine translation of scientific abstracts and terminologies has the potential to support health professionals and biomedical researchers in some of their activities. In the fifth edition of the WMT Biomedical Task, we addressed a total of eight language pairs. Five language pairs were previously addressed in past editions of the shared task, namely, English/German, English/French, English/Spanish, English/Portuguese, and English/Chinese. Three additional languages pairs were also introduced this year: English/Russian, English/Italian, and English/Basque. The task addressed the evaluation of both scientific abstracts (all language pairs) and terminologies (English/Basque only). We received submissions from a total of 20 teams. For recurring language pairs, we observed an improvement in the translations in terms of automatic scores and qualitative evaluations, compared to previous years.

[*] The author list is alphabetical and does not reflect the respective author contributions.

1 Introduction

Automatic translation aims to alleviate the language barrier by providing access to information for readers not familiar with the original language used to write documents. Access to accurate biomedical information is specifically critical and machine translation (MT) can contribute to making health information available to health professionals and the general public in their own language. It can also contribute to biomedical research by assisting with the writing of research reports in English. In addition, machine translation can provide the opportunity to enhance the use of natural language processing (NLP) tools and methods for low-resource languages by the development of resources through translation or by making tools available through text translation into resource rich languages.

Herein, we describe the fifth edition of the WMT Biomedical task,[1] which aims to evaluate the auto-

[1] http://www.statmt.org/wmt20/

matic translation of a variety of biomedical texts.

The first edition of the task (Bojar et al., 2016) focused on biomedical scientific abstracts in three language pairs. The second edition of the task offered ten language pairs and addressed scientific abstracts as well as patient-oriented health information (Jimeno Yepes et al., 2017). The third edition of the task offered six language pairs and addressed scientific abstracts (Neves et al., 2018). The fourth edition of the task offered ten language pairs. It addressed scientific abstracts and introduced the task of terminology translation (Bawden et al., 2019). This year's edition of the task continues to address the translation of scientific abstracts and terminologies. It builds on previous tasks by offering a large range of training and test sets to support participants' systems. The following language pairs are 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 (en2de)
- 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)

Similar to previous years, our test sets consist of scientific abstracts retrieved from the MEDLINE® database. In continuation with last year's task (Bawden et al., 2019), we also provide a test set for the automatic translation of biomedical terminologies. Below, we highlight some new aspects introduced in the 2020 edition of the shared task:

- We address three new language pairs, namely, en/eu, en/it, en/ru[2].
- We include a novel test set for the automatic translation of biomedical terminologies from English to Basque (cf. Section 2.2.1)
- During the construction of the test sets, and after the manual validation of the automatic alignment, we ran a pilot project for a couple of languages in which we manually fine-tuned the alignment of the test sets (cf. Section 2.2.3).
- We ran a second pilot study in which we split the sentences according to the reported original language of the abstract (cf. 2.2.3).
- Three of our tests sets, namely, de/en, ru/en and zh/en, were included as test suites in the WMT News Task (cf. Section 5.2).
- Participants were asked to provide details about their systems through an online survey (cf. Tables 6, 7, and 9).
- Our manual validation included whole abstracts, in addition to (correctly aligned) sentence pairs (cf. Section 6.1).
- We ran a third pilot study in which two experts validated submissions for certain language pairs, in which one was a native speaker of the source language, while the other a native speaker of the target language (cf. Tables 17 and Table 20).
- Our methodology for ranking the systems based on the manual validation considered a significance test and a points-based schema (cf. Section 6.1).

This article is structured as follows: Section 2 presents the details of the generation of our training and test sets, for both the scientific abstracts and the terminology, as well as manual validation of the quality of the test sets. Section 3 describes our baseline systems, which are used as comparison in the automatic evaluation. We list all teams that participated in our task in Section 4, as well as details of the methods behind their systems and the in-domain and out-of-domain data that was used. The results of the automatic evaluation based on the BLEU and chrF scores are presented in Section 5, while the ones for the manual evaluation are presented in Section 6. Finally, we discuss various topics related to the shared task in Section 7.

biomedical-translation-task.html

[2]Throughout the paper, we will refer to en/ru (or ru/en), for instance, when referring to the language pair in general, without specifying the translation direction. When making reference to the direction, we will use either en2ru or ru2en, for instance.

2 Training and test data

We provided training data of MEDLINE abstracts for it/en and ru/en, since training data for some of the other languages was already available from previous years. As for the tests sets, we released test sets for scientific abstracts and for terminologies, as summarized below:

- 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)
  - Spanish/English (both directions)
- Terms from biomedical terminologies:
  - English to Basque

Additional details are presented in Table 1. In this section we describe the details about the construction of resources that we released for the shared task.

2.1 Training data

We released training data from MEDLINE for two of the new language pairs that we address this year, namely, English/Italian and English/Russian.

We relied on the latest version of the MEDLINE baseline[3] available at the time of data preparation. We retrieved all the abstracts that were available in Italian and English, or in Russian and English. We summarize below the steps that we followed to process the data:

1. Abstracts were parsed using the `pubmed_parser` library.[4]
2. The language of these abstracts, as identified by MEDLINE meta-data, was confirmed with the `langdetect` library.[5]
3. Sentences in the abstracts were split using the `syntok` library.[6]
4. These sentences were automatically aligned using the `GMA` tool[7] using specific stopword lists for each language.

We obtained a total of 1,675 parallel documents for it/en and 6,029 for ru/en. The training data is available in our GitHub repository.[8]

In regard to English-Basque scientific abstract translation, we could not release any in-domain parallel data, as very little is still written in Basque in the medical domain. However, we provided other corpora that can help with training machine translation models. These include out-of-domain parallel corpora such as the TED talks,[9] the datasets available on the OPUS repository[10] and the WMT16 IT translation shared-task.[11] Additionally, we released in-domain monolingual corpora[12] that include translations of examples of hospital notes, automatic translations of SNOMED CT terms (Perez-de Viñaspre and Oronoz, 2015), and medical domain articles from Wikipedia. Finally, we released a recent dump of the whole Wikipedia (01/2020) as a large, out-of-domain monolingual corpus.[13]

For the terminology translation task, on behalf of Osakidetza (Basque Public Health System), we released 27,900 terms of the Basque ICD-10-CM. These descriptions were manually validated by the institution's translation team. 25,900 descriptions where released as a training set, keeping the remaining 2,000 for the development set. Both sets are plain text, and they have not been tokenized. On average, in the training set, each term comprises 6.72 words (split on whitespace and punctuation), 1 being the minimum and 27 the maximum. For the development set, the average word count is 6.75, 1 being the minimum and 25 the maximum.

2.2 Test sets

All test sets were released on June 29th, 2020 and the participants could submit results until July 9th, 2020. The test sets for de/en, ru/en and zh/en were

[3] https://www.nlm.nih.gov/databases/download/pubmed_MEDLINE.html released at the end of 2019.

[4] https://github.com/titipata/pubmed_parser

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

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

[7] https://nlp.cs.nyu.edu/GMA/

[8] https://github.com/biomedical-translation-corpora/corpora

[9] https://wit3.fbk.eu/mt.php?release=2018-01

[10] http://opus.nlpl.eu/

[11] http://www.statmt.org/wmt16/it-translation-task.html

[12] https://drive.google.com/drive/u/2/folders/1cQmiywDRcAeHeRuZfaF-zuoG7DQHO4CQ

[13] https://drive.google.com/drive/u/2/folders/1BjScNNvMbVOzrD3KWA0D0UGR33j6Lg83

Language pairs	MEDLINE training		Abstracts test		Terminology test
	Documents	Sentences	Documents	Sentences	Terms
en2eu	-	-	40	375	2,000
de2en	-	-	50	612/652	-
en2de			50	783/742	-
es2en	-	-	50	533/629	-
en2es			50	618/562	-
fr2en	-	-	50	563/584	-
en2fr			50	757/731	-
it2en	1,675	15,950/ (it)	50	549//716	-
en2it		20,615 (en)	50	624/468	-
pt2en	-	-	50	498/637	-
en2pt			50	544/466	-
ru2en	6,029	52,544/ (ru)	50	463/523	-
en2ru		61,494 (en)	50	553/484	-
zh2en	-	-	50	412/622	-
en2zh			50	514/343	-

Table 1: Number of documents, sentences and terms in the training and test sets released for this shared task.

also included as test suites of the WMT news task and released on June 22nd, 2020. In the following we describe details of the test set construction.

2.2.1 Terminology

In addition to the training set of ICD-10-CM Basque terms, there were 2,000 more terms for the test set. Again, this set was not tokenized. On average, each term comprises 7.74 words, 1 being the minimum word count and 25 the maximum. Unfortunately, at the time of releasing the test set, due to a confusion on behalf of the organizers, the development set was provided as test for all participants, and was used for evaluation. The planned test set has been publicly released for download.[14]

2.2.2 Basque abstracts

The Basque language appears in MEDLINE as a subject of study but not systematically as a writing language, so there is not a sufficient corpus for training in Basque in MEDLINE. The abstracts used in the test are taken from the journal Osagaiz,[15] the first journal on medicine written entirely in Basque (with abstracts also in English).

Osagaiz[16] was published for the first time in 2017 and every year it publishes a volume with at least two numbers. Its main objective is to be a way of communicating the scientific findings of the Basque health community in Basque. Three volumes have been used in the test (years 2017, 2018 and 2019); that is, 6 numbers with 40 abstracts in both English and Basque. The Basque abstracts dataset consists of 375 sentences (8,651 tokens in English with 23.07 tokens per sentence, and 7459 tokens in Basque with 19.89 tokens per sentence).

2.2.3 MEDLINE abstracts

We followed a similar approach to the one we used in previous years. However, we carried out two novel pilot studies this year: (a) a manual improvement of the alignment after the manual validation, and (b) a selective split of the abstracts for the translation directions based on the original language of the abstract.

For the test sets, we retrieved the citations that were published in 2020 and were not included in any of the previously released training and test sets. We parsed the articles and checked the language using the same tools as described for the training data above. We split the sentences for all languages using the `syntok` library, except for zh/en where it was sufficient to split sentences according to the Chinese punctuation (。) that marks the end of a sentence. Sentence alignment was carried out for all languages (except for zh/en) with the `GMA` tool using specific stopword lists for each language. For zh/en, we used the `Champollion` tool[17] with the same configurations and stopword lists since 2018.

[14] https://drive.google.com/drive/folders/1KXUjEBUzudi81y5rxm33UxkmRY9RSKMj

[15] http://www.osagaiz.eus/

[16] The contents from Osagaiz are licensed under Creative Common Attribution-ShareAlike 3.0 unported (CC BY-SA 3.0) https://creativecommons.org/licenses/by-sa/3.0/deed.en

[17] http://champollion.sourceforge.net/

We randomly retrieved a set of 100 abstracts for each language pair, and the automatic aligned sentences were manually validated by native speakers of the foreign languages using the Appraise tool (Federmann, 2010). Results of the validation are shown in Table 2. For the ru/en set, an additional set of 100 abstracts were randomly retrieved for a second round of manual validation. This was due to the low quality of the alignments that we obtained in the first round of validation. The official test set for ru/en was composed of the abstracts with better quality from the totality of 200 abstracts that were validated.

As a pilot study this year, we performed a manual correction of the alignment which were identified as not being correct during the validation in the Appraise tool. This step was only carried our for the es/en, fr/en, ru/en, and zh/en test sets. For all these languages, this extra step increased alignment quality (cf. Table 2): from 80.54% correctly aligned sentences to 91.49% for fr/en, from 55.27% to 61.96% for ru/en, from 83.57% to 88.07% for es/en, and a slight improvement from 63.84% to 64.43% for zh/en.

Most of the remaining sentences are in fact titles in English, for which a translation in the foreign language is not available from MEDLINE. For zh/en, the manual corrections addressed mismatching sentence splitting policies for abstract subsections such as *OBJECTIVE: To investigate...* and *METHODS: We used xyz...* The GMA tool split such a text into two sentences, but the Champollion tool kept it as one sentence. With this extra step, affected sentences that were marked as "NO_ALIGNMENT" became "TARGET_GREATER_SOURCE" (cf. Table 2 for the alignment categories).

Finally, the set of 100 abstracts was randomly split into two sets of 50 abstracts, for each translation direction, e.g., es2en and en2es. Exception was made for the fr/en test set. Following the recommendations of Graham et al. (2019), we tried to split the data sets depending on which language we hypothesized was the abstract's source language. For articles with a documented "TT" field (vernacular, i.e. French, title) in the MEDLINE citation, we considered that the source language was French and otherwise, English. As a result, the en/fr test only contains abstract originally written in English. However, since only 20 abstract in our set were originally written in French, the fr/en set still contains a mix of source languages. This suggests that vernacular titles should be considered in the initial set selection.

3 Baselines

We provided our baseline systems for all language pairs in the scientific abstracts translation subtask.

There were two categories of baseline: for en/zh, en/fr, en/de, en/pt and en/es the models used for each direction were transformers (Vaswani et al., 2017) trained by us using MarianNMT (Junczys-Dowmunt et al., 2018) with the following settings: joint BPE of 40,000, beam size 16. These parameters were chosen by tuning on a single direction of a single language pair: English to German. Each of the 10 models were trained for up to two days. The training was stopped when there were no improvements on the validation dataset for more than 10 epochs, as measured through cross-validation score. The corpora we used to train the models were the same as last year – when we had baselines generated using RNN-based sequence2sequence models: the UFAL medical corpus (UFA) without the "Subtitles" subset, and as validation we again used Khreshmoi (Dušek et al., 2017).

For en/it and en/ru and en/eu we used the Helsinki-NLP/opus-mt-SRC-TRG models (Tiedemann and Thottingal, 2020) included in the huggingface transformers library [18], trained with MarianNMT on the entirety of the OPUS corpora (Tiedemann, 2012). These models are not uniformly good; they performed very well for Italian, but fairly poor for Russian and Basque.

Discussion. It is interesting that the models for English to/from Italian performed so well in the biomedical task, as they were trained on generic text, not targeting the biomedical domain. It is interesting in general to what extent models that excel on generic text (e.g. news) perform well on the biomedical texts as well.

4 Teams and systems

This year, 22 teams submitted a total of 151 runs. Two teams withdrew after submitting their runs. The remaining teams were from China (7 teams), Spain (3 teams), France (2 teams), the United Kingdom (2 teams), Armenia (1 team), Australia (1 team), Brazil (1 team), India (1 team), Ireland (1 team) and Pakistan (1 team). Table 3 presents the

[18] https://huggingface.co/transformers/

Language	OK	Source>Target	Target>Source	Overlap	No Align.	Total
de/en	909 (70.85%)	63 (4.91%)	104 (8.11%)	52 (4.05%)	155 (12.08%)	1,283
es/en	931 (83.57%)	29 (2.60%)	54 (4.85%)	7 (0.63%)	93 (8.35%)	1,114
es/en §	1,026 (88.07%)	9 (0.78%)	4 (0.34%)	0 (0%)	126 (10.82%)	1,165
fr/en	985 (80.54%)	34 (2.78%)	74 (6.05%)	6 (0.49%)	124 (10.14%)	1,223
fr/en §	1225 (91.49%)	7 (0.52%)	8 (0.60%)	2 (0.15%)	97 (7.24%)	1,339
it/en	636 (60.40%)	51 (4.84%)	150 (14.25%)	60 (5.70%)	156 (14.81%)	1,053
pt/en	799 (78.41%)	37 (3.63%)	66 (6.48%)	20 (1.96%)	97 (9.52%)	1,019
ru/en *	947 (53.14%)	67 (3.76%)	186 (10.44%)	65 (3.65%)	517 (29.01%)	1,782
ru/en **	472 (55.27%)	33 (3.86%)	94 (11.01%)	32 (3.75%)	223 (26.11%)	854
ru/en §	562 (61.96%)	30 (3.3%)	60 (6.61%)	28 (3.09%)	228 (25.14%)	908
zh/en	535 (63.84%)	36 (4.30%)	135 (16.11%)	9 (1.07%)	123 (14.68%)	838
zh/en §	540 (64.43%)	137 (16.35%)	142 (16.95%)	9 (1.07%)	10 (1.19%)	838

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 for the totality (200 abstracts) for ru/en. ** Results for the selected test set (100 abstracts) for ru/en. § Results after manual correction of sentence segmentation and/or alignment.

list of teams that submitted at least one run to the biomedical task.

At least one run was submitted for each language pair offered, with the most runs submitted for English to Basque (terminology test set, 24 runs) and English to Chinese (MEDLINE test set, 18 runs). 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.

During the automatic evaluation, we observed that some teams obtained extremely high BLEU scores, which were close to 0.9. Those teams had trained their systems on the MEDLINE database, and the training data potentially included our test sets. Unfortunately, as opposed to previous years, we forgot to inform participants on our website that this practice was not allowed. Therefore, we offered the opportunity for these teams to re-submit their runs, but without training on MEDLINE. The Wei-Bot team was the only one to submit new runs.

In an effort to increase the level of detail in the system description and the comparability between systems, we asked participants to fill in a survey with key information regarding the translation method used, as well as the in-domain and general datasets used for training. The survey comprised 14 questions covering the translation methods and corpora used. Teams indicated their primary submission, which was considered for manual evaluation. On average, submission time for one language pair was 6 minutes and 28 seconds (Median: 3 minutes and 35 seconds). All teams used transformer-based neural machine translation (except for team TRAMECAT, who used sequence2sequence) and mostly relied on existing implementations: 19 teams submitted runs using available libraries, one team submitted runs using a mix of libraries and in-house implementations, one team submitted runs exclusively relying on their own implementation of NMT. Teams often used the same setup for a range of language pairs. Table 6 shows details about the teams methods.

For in-domain data, teams used the training data distributed by us and many of the sources described in (Névéol et al., 2018). Tables 7 and 8 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. Interestingly, some teams used similarity measures based on biomedical corpora to extract additional biomedical sentences from out-of-domain corpora. Out-of-domain data was also used in the form of pre-trained base models. Table 9 shows details of the out-of-domain data used by the teams.

5 Automatic evaluation

Following (Mathur et al., 2020), we used chrF (Popović, 2015) as well as BLEU (Papineni et al., 2002) as automatic metrics. chrF scores are obtained using the `nltk` implementation.[19]

5.1 MEDLINE

Similarly to previous years, we compared the submitted translations to the reference translations

[19] https://www.nltk.org/_modules/nltk/translate/chrf_score.html

Team ID	Institution
ADAPT (Nayak et al., 2020)	Dublin City University, Ireland
ai_not_intellegent	ai_not_intellegent, China
Alibuba	Alibab DAMO Academy, China
baidu_translation	Baidu translation, China
Elhuyar_NLP (Corral and Saralegi, 2020)	Elhuyar Foundation, Spain
Huawei United (Peng et al., 2020)	Huawei Technologies, China
Ixamed (Soto et al., 2020)	University of the Basque Country, Spain
LIMSI (Abdul Rauf et al., 2020)	LIMSI-CNRS, France
NLE	Naver Labs Europe, France
nrpu-fjwu (Naz et al., 2020)	Fatima Jinnah Women University, Pakistan
one_connect_000	OneConnect AI Lab, China
OOM_20	Atman Tech, India
Sheffield (Soares and Vaz, 2020)	University of Sheffield, UK
TMT (Wang et al., 2020)	Tencent AI Lab, China
TRAMECAT	Universitat Oberta de Catalunya, Spain
UNICAM (Saunders and Byrne, 2020)	University of Cambridge, UK
UNICAMP_DL (Lopes et al., 2020)	University of Campinas, Brazil
UTS_NLP (Jauregi Unanue and Piccardi, 2020)	University of Technology Sydney, Australia
Wei-Bot	East China Normal University, China
YerevaNN (Hambardzumyan et al., 2020)	YerevaNN, Armenia

Table 3: List of the participating teams.

Teams	en2eu	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh	Total
ADAPT	A3T3	-	-	-	-	-	-	-	6
ai_not_intellegent	-	-	-	-	-	-	-	A3	3
Alibuba	-	-	-	-	-	-	-	A1	1
baidu_translation	-	-	-	-	-	-	-	A1	1
Elhuyar_NLP	A3T3	-	A3	-	-	-	-	-	9
Huawei United	-	A3	-	A2	A2	-	A2	A3	12
Ixamed	A3T3	-	A3	-	-	-	-	-	9
LIMSI	-	-	-	A2	-	-	-	-	2
NLE	-	A3	-	-	-	-	-	-	3
nrpu-fjwu	-	-	-	A1	-	-	-	-	1
one_connect_000	-	-	-	-	-	-	-	A1	1
OOM	-	-	-	-	-	-	-	A2	2
Sheffield	-	-	A1	A1	A1	A1	A1	-	5
TMT	-	A3	-	-	-	-	-	A3	6
TRAMECAT	-	-	A1	A1	-	-	A1	A1	4
UNICAM	-	A3	A3	-	-	-	-	-	6
UNICAMP	-	-	-	-	-	A2	-	-	2
UTS_NLP	A3T3	-	-	-	-	-	-	-	6
Wei-Bot	-	-	-	-	-	-	-	A2	2
YerevaNN	-	A2	-	-	-	-	A3	-	5
Total	24	14	11	7	3	3	7	17	86

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.

using BLEU with the `MULTI-EVAL v14` tool[20] provided by the Moses package (Koehn et al., 2007). This means as well that we reused the tokenization approach used for Chinese. Results for MEDLINE BLEU are shown in Tables 10 and 11.

[20] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v14.pl

5.2 News

The test set of our challenge was included in the News challenge data set. We identified the translations in the News files and used the same evaluation procedure as applied to MEDLINE abstracts. Results of the systems are shown in Tables 12 and 13.

Teams	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en	Total
ai_not_intellegent	-	-	-	-	-	-	A3	3
Alibuba	-	-	-	-	-	-	A1	1
baidu_translation	-	-	-	-	-	-	A1	1
Huawei United	A3	-	A2	A2	-	A2	A2	11
Ixamed	-	A3	-	-	-	-	-	3
NLE	A3	A1	A1	A1	-	-	-	6
nrpu-fjwu	-	-	A3	-	-	-	-	3
one_connect_000	-	-	-	-	-	-	A1	1
OOM	-	-	-	-	-	-	A2	2
Sheffield	-	A1	A1	A1	A1	A1	-	5
TMT	A3	-	-	-	-	-	A1	4
TRAMECAT	-	A1	A1	-	-	A1	A1	4
UNICAM	A3	A3	-	-	-	-	-	6
UNICAMP	-	-	-	-	A2	-	-	2
Wei-Bot	-	-	-	-	-	-	A2	2
YerevaNN	A3	-	-	-	-	A2	-	5
Total	15	9	8	4	3	6	14	59

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
ADAPT	all	Marian NMT	Yes	No	Yes	No
ai_not_intellegent	zh2en	Fairseq	Yes	Yes	No	No
ai_not_intellegent	en2zh	Own	No	Yes	No	MASS
Alibuba	zh2en	OpenNMT	Yes	No	Yes	transformer-base
Alibuba	en2zh	OpenNMT	No	Yes	Yes	transformer-base
baidu_translation	all	paddle	Yes	No	Yes	paddle
Elhuyar_NLP	all	OpenNMT	Yes	No	en2eu	No
Huawei United	en/de	Own	Yes	No	No	FB-PLM
Huawei United	all but en/de	Own	Yes	No	zh2en	No
Ixamed	all	Open NMT	Yes	No	No	No
LIMSI	all	Fairseq	Yes	Yes	Yes	Yes
NLE	de2en	Fairseq	Yes	No	Yes	No
NLE	fr2en	Fairseq	Yes	Yes	Yes	No
NLE	it2en	Fairseq	Yes	Yes	Yes	No
nrpu-fjwu	all	Fairseq	Yes	No	Yes	fr2en
OOM_20	all	tensor2tensor, modified	Yes	Yes	-	-
Sheffield	all but ru/en	Tensorflow	Yes	No	{es,fr,it,pt}2en	No
Sheffield	ru2en, en2ru	Tensorflow	Yes	Yes	ru2en	No
TMT	all	Fairseq	Yes	No	Yes	No
TRAMECAT	all	MarianNMT	Yes	No	No	No
UNICAM	all	Tensor2Tensor	No	Yes	No	No
UNICAMP_DL	all	T5, Huggingface	No	Yes	No	T5 HuggingFace
UTS_NLP	all	Fairseq, BERT-NMT	Yes	No	Yes	Yes
Wei-Bot	all	Fairseq	Yes	No	Yes	MASS
YerevaNN	all	Fairseq?	No	Yes	ru2en	XLM-R

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

5.3 Basque abstracts

For the Basque abstract we used the same evaluation tool as for MEDLINE (MULTI-EVAL), and the results are presented in Table 14.

5.4 Terminology

For the evaluation of terminology we provide two metrics for the en2eu task: (i) accuracy, by relying on strict matches (case-insensitive) between ground truth and predictions; and (ii) sentence-level BLEU score, as measured by the `nltk` module sentence-BLEU.[21] Results are presented in Table 15.

[21] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

Language pair	team	Parallel corpus	size (sentence pairs)	Monolingual corpus	size (sentences)
en/de	Huawei	MEDLINE abstracts corpus supplied by organizers.	29 k	No	-
	NLE	MEDLINE abstracts corpus supplied by organizers.	34,710	No	-
	UNICAM	TRAINING: UFAL medical and MEDLINE abstracts corpus supplied by organizers. FINE-TUNING: MEDLINE abstracts	TRAINING: 2.2M FINE-TUNING: 28K	No	-
	TMT	UFAL medical and MEDLINE abstracts corpus supplied by organizers.	2.5M	UFAL (en)	5.4M
	Yereva_NN	MEDLINE abstracts corpus supplied by organizers; alignment was fixed using XLM-R	32,466	No	-
en/es	Elhuyar_NLP	Scielo and corpora supplied by organizers.	560k	No	-
	Ixamed	MEDLINE corpus supplied by organizers and TAUS Corona Crisis Corpus	1,290,201	No	-
	UNICAM	TRAINING: UFAL medical, Scielo (Neves et al., 2016), and MEDLINE abstracts corpus supplied by organizers. FINE-TUNING: MEDLINE abstracts	TRAINING: 1.3M FINE-TUNING: 67K	No	-
	Sheffield	BVS, EMEA, Scielo (Soares et al., 2018) and MEDLINE corpus supplied by organizers as well as new crawled PubMed data. The data was checked against the official test set to avoid including test data during training.	2.5M	No	-
	TRAMECAT	Biomedical translation repository, EMEA, IBECS, ICD10, Kreshmoi, MEDLINE corpus supplied by organizers, in-house MEDLINE (dated 2018), Medem glossaries, MSDManuals, Portal Clinic corpus, Scielo, SNOMED	7,232,784	No	-
en/eu	ADAPT	Data provided by the organisers	-	Common Crawl selected by TermFinder	200k (en) 41,151 (eu)
	Elhuyar_NLP	WMT20 shared task bilingual training data, internal medical corpus, and synthetically generated data from the WMT19 EN-ES shared task	Around 350k segments	SNOMED descriptions, hospital notes and wikipedia medical articles (en)	Around 110k segments
	Ixamed	-	-	-	-
	UTS_NLP	ICD-10 codes translations	25900	SNOMED terms, hospital notes and wikipedia medical articles (en)	total of 60,000 sentences)
en/fr	Huawei	MEDLINE abstracts corpus supplied by organizers, in-domain lexicon	4M bitext, 59k lexicon	Yes (en)	22M
	LIMSI	Cochrane, Taus and corpora supplied by organizers	3,951,013	LISSA (Griffon et al., 2017) (fr)	395,699
	NLE	In-domain parallel data obtained from WMT and OPUS	-	No	-
	nrpu-fjwu	Corpora supplied by organizers (MEDLINE, Scielo, EDP, UFAL).	3,408,327	No	-
	Sheffield	EMEA and MEDLINE corpus supplied by organizers as well as new crawled PubMed data. Prior to training, the data was checked against the official test set to avoid including test data during training.	3.42M	MEDLINE (en)	2M
	TRAMECAT	EMEA, MEDLINE corpus supplied by organizers, PatTR medical, Scielo(Neves et al., 2016)	4,2 M	No	-
en/it	Huawei	MEDLINE abstracts corpus supplied by organizers.	219k	No	-
	NLE	MEDLINE corpus supplied by organizers, TAUS Corona Corpus, OPUS	-	No	-
	Sheffield	EMEA and MEDLINE corpus supplied by organizers as well as new crawled PubMed data. The data was checked against the official test set to avoid including test data during training.	1.0M	MEDLINE (en)	1M
en/pt	Sheffield	BVS, EMEA, Scielo (Soares et al., 2018) and MEDLINE corpus supplied by organizers as well as new crawled PubMed data. The data was checked against the official test set to avoid including test data during training.	5.5M	MEDLINE (en)	2M
	UNICAMP_DL	EMEA corpus, MEDLINE corpus supplied by organizers, Scielo (Soares et al., 2018), a corpus of theses and dissertations abstracts (BDTD) from CAPES, JRCAcquis.	6,606,858	MEDLINE (en)	2M

Table 7: Overview of in-domain corpora used by participating teams. Information is self reported through our survey for each selected "best run".

Language pair	team	Parallel corpus	size (sentence pairs)	Monolingual corpus	size (sentences)
en/ru	Huawei	MEDLINE abstracts corpus supplied by organizers.	32 k	No	-
	Sheffield	MEDLINE corpus supplied by organizers as well as new crawled PubMed data. The data was checked against the official test set to avoid including test data during training.	15k	MEDLINE (en)	100k
	TRAMECAT	MEDLINE corpus supplied by organizers, Corona TAUS corpus, ICD10 (subset)	240,998	No	-
	Yereva_NN	MEDLINE abstracts corpus supplied by organizers; alignment was fixed using XLM-R	37,201	No	-
en/zh	ai_not_intel...	Web crawl augmented by back translation	"3.G in text"	Yes	
	Alibuba	PubMed articles in Chinese	"1.2G text"	No	-
	Baidu	"inhouse dataset"	"12.5G text"	No	-
	Huawei	in-domain lexicon	59k	Yes (en)	62M
	OOM_20	Abstracts from Chinese medical papers	3 M	medical papers	(zh) 10M, (en) 20M
	TMT	No	-	Yes (en)	5.4M
	TRAMECAT	Corona TAUS corpus	450,507	No	-
	Wei-Bot	Pubmed Crawl	3M	Wikipedia (en, zh)	-

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

6 Manual evaluation

We manually validated a sample for each primary run in order to compare the performance between teams as well as to the reference translations. In this section we present details of the evaluation and results that we obtained.

6.1 MEDLINE abstracts

Similarly to previous years, we aimed to validate a total of 100 sampled sentences per primary run. This year, we manually validated not only single sentences, but also whole abstracts. The selection of abstracts to be validated for each language pair followed the procedure described below:

1. Randomly select an abstract.
2. Check whether the percentage of perfectly aligned sentences is at least of 80%.
3. Retrieve all perfectly (i.e., OK) aligned sentences from the abstract.
4. Repeat steps 1 to 3 above if the total number of selected sentences (over all selected abstracts) is below 100.

In the case of zh2en and en2zh, due to the large number of submissions that we received, the manual validation was restricted to the abstracts. However, these were selected using the same approach described above. In addition, one team re-submitted their results after the official test period, and we note that these re-submissions are not fully comparable to the ones submitted before the period (see Tables 10, 11 and 22).

Due to time constraints, we could not validate all planned abstracts and sentences that were selected for de2en, but only about half of them. Further, and due to the same reason, the validation for es2en and pt2en was limited to a few abstracts (and its sentences) and was validated as a collaboration between two experts: (1) one who was a native speaker of the source language and who checked whether any information that was included in the source text was missing in the translation; and (2) one who was a native speaker of English, and who was in charge of checking the quality of the English translations.

If the information about the primary run was not available for a particular team and test set, we considered the run with the highest BLEU score. We only considered for manual validation those teams that provided detailed information about their system by filling out a survey mentioned in Section 4. The runs that we considered are listed below:

- en2de (5 teams): Huawei United (run3), NLE (run3), TMT (run1), UNICAM (run3), YerevaNN (run3)
- en2es (5 teams): Elhuyar (run1), Ixamed (run1), Sheffield (run1), TRAMECAT (run1), UNICAM (run3)
- en2fr (5 teams): Huawei United (run2), LIMSI (run1), Sheffield (run1), TRAMECAT (run1), nrpu-fjwu (run1)

Language pair	team	Parallel corpus	size (sentence pairs)	Monolingual corpus	size (sentences)
en/de	Huawei	TRAINING: in-house bitext FINE-TUNING: tfidf filtering of training corpus	TRAINING: 2.3M FINE-TUNING: 27K	Yes (de)	2.3M
	NLE	All de-en parallel data supplied by WMT20 News Task	44.8M	NewsCrawl	269M (en) 440M (de)
	UNICAM	For pre-training, corpus supplied by the WMT 2018 news task organizers	17M	No	-
	TMT	Corpus supplied by the WMT 2020 News task organizers	37.8M	No	-
	Yereva_NN	No OOD data was used directly, but the base models we had fine-tuned were trained on news data (Ng et al., 2019)	-	No	-
en/es	Elhuyar_NLP	Paracrawl v5 corpus	33M	No	-
	Ixamed	No	-	No	-
	UNICAM	No	-	No	-
	Sheffield	No	-	No	-
	TRAMECAT	UNPC parallel corpus: segments selected by similarity (using a language model on the English part)	5M	No	-
en/eu	ADAPT	Data provided by the organisers	-	CommonCrawl (eu)	400K
	Elhuyar_NLP	Synthetic data was obtained by backtranslating an internal ES-EU corpus from Spanish to English	Around 7M segments	No	-
	Ixamed	-	-	-	-
	UTS_NLP	Out of domain parallel corpora provided by WMT2020 biomedical translation organizers.	approx. 0.6M	Wikipedia (eu)	1.5M
en/fr	Huawei	news and other data (in-house)	123M	Yes (en)	62M
	LIMSI	No	-	No	-
	NLE	OOD WMT and OPUS	-	Back Translation en2ko	8M
	nrpu-fjwu	Medical domain sentences retrieved from books, news commentary and wikiPedia parallel corpus.	243,182	medical sentences retrieved from wikiPedia (fr)	-
	Sheffield	No	-	No	-
	TRAMECAT	UNPC parallel corpus: segments selected by similarity (using a language model on the English part)	5M	No	-
en/it	Huawei	in-house general domain data like news	150M	No	-
	NLE	Paracrawl, OPUS, UN Political corpus	-	English sentences back-translated	9.2M
	Sheffield	No	-	No	-
en/pt	Sheffield	No	-	No	-
	UNICAMP_DL	ParaCrawl dataset (subset)	5M	No	-
en/ru	Huawei	No	-	No	-
	Sheffield	ParaPat corpus of Patents (Soares et al., 2020)	4.3M	MEDLINE (en)	100k
	TRAMECAT	UNPC parallel corpus: segments selected by similarity (using a language model on the English part)	5M	No	-
	Yereva_NN	No OOD data was used directly, but the base models we had fine-tuned were trained on news data (Ng et al., 2019)	-	No	-
en/zh	ai_not_intel...	Corpus supplied by the WMT 2020 News task organizers	"3.G in text"	No	-
	Alibuba	No	-	No	-
	Baidu	No	-	No	-
	Huawei	"inhouse dataset"	186M	No	-
	OOM_20	Corpus supplied by the WMT 2020 News task organizers	10 M	No	-
	TMT	No	-	Yes (en)	5.4M
	TRAMECAT	UNPC parallel corpus: segments selected by similarity (using a language model on the English part)	5M	No	-
	Wei-Bot	No	-	No	-

Table 9: Overview of out-of-domain (OOD) corpora used by participating teams. Information is self reported through our survey for each selected "best run".

- en2it (2 teams): Huawei United (run2), Sheffield (run1)
- en2pt (2 teams): Sheffield (run1), UNICAMP_DL (run1)
- en2ru (4 teams): Huawei United (run2), Sheffield (run1), TRAMECAT (run1), YerevaNN (run3)
- en2zh (8 teams): ai_not_intellegent (run1),

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh
Alibuba	Run1	-	-	-	-	-	-	0.3346*
Elhuyar_NLP	Run1	-	0.4498*	-	-	-	-	-
	Run2	-	0.4493	-	-	-	-	-
	Run3	-	0.4394	-	-	-	-	-
Huawei_United	Run1	0.3317	-	0.4351*	0.4257	-	0.3464	0.4378
	Run2	0.362	-	0.4351*	0.4257*	-	0.3464*	0.4546
	Run3	0.3689*	-	-	-	-	-	0.4378*
Ixamed	Run1	-	0.4171*	-	-	-	-	-
	Run2	-	0.3836	-	-	-	-	-
	Run3	-	0.3858	-	-	-	-	-
LIMSI	Run1	-	-	0.3837*	-	-	-	-
	Run2	-	-	0.3673	-	-	-	-
	Run3	-	-	0.2564	-	-	-	-
NLE	Run1	0.3641	-	-	-	-	-	-
	Run3	0.3394	-	-	-	-	-	-
	Run3	0.3562*	-	-	-	-	-	-
OOM_20	Run1	-	-	-	-	-	-	0.4686*
	Run2	-	-	-	-	-	-	0.4633*
Sheffield	Run1	-	0.4493*	0.3049*	0.2073*	0.4744*	0.2573*	-
TMT	Run1	0.3524*	-	-	-	-	-	0.3943*
	Run2	0.3495	-	-	-	-	-	-
	Run3	0.3457	-	-	-	-	-	-
TRAMECAT	Run1	-	0.4361*	0.3489*	-	-	0.2661*	0.2725*
UNICAMP_DL	Run1	-	-	-	-	0.4095*	-	-
	Run2	-	-	-	-	0.3660	-	-
UNICAM	Run1	0.3288	0.4572	-	-	-	-	-
	Run2	0.3282	0.4672	-	-	-	-	-
	Run3	0.3318*	0.4662*	-	-	-	-	-
Wei-Bot	Run1	-	-	-	-	-	-	0.5557*§
	Run2	-	-	-	-	-	-	0.5169§
YerevaNN	Run1	0.3517	-	-	-	-	0.3263	-
	Run2	-	-	-	-	-	0.3936	-
	Run3	0.3520*	-	-	-	-	0.3787*	-
ai_not_intellegent	Run1	-	-	-	-	-	-	0.4462
	Run2	-	-	-	-	-	-	0.4148
	Run3	-	-	-	-	-	-	0.4225
baidu_translation	Run1	-	-	-	-	-	-	0.3400
nrpu-fjwu	Run1	-	-	0.3572*	-	-	-	-
one_connect_000	Run1	-	-	-	-	-	-	0.3125*
Baseline	-	0.2845	0.3813	0.3345	0.3954	0.4149	0.2259	0.2319

Table 10: BLEU scores for “OK" aligned test sentences, from English. * Indicates the primary run as indicated by the participants. § Runs submitted after the official test period.

Alibuba (run1), baidu_translation (run1), Huawei United (run3), OOM_20 (run1), TMT (run1), TRAMECAT (run1), Wei-Bot (run1)

- de2en (5 teams): Huawei United (run3), NLE (run3), TMT (run1), UNICAM (run3), YerevaNN (run3)
- es2en (4 teams): Ixamed (run1), Sheffield (run1), TRAMECAT (run1), UNICAM (run3)
- fr2en (5 teams): Huawei United (run2), NLE (run1), Sheffield 8run1), TRAMECAT (run1), nrpu-fjwu (run1)
- it2en (3 teams): Huawei United (run2), Sheffield (run1), NLE (run1)
- pt2en (2 teams): Sheffield (run1), UNICAMP_DL (run1)
- ru2en (4 teams): Huawei United (run2), Sheffield (run1), TRAMECAT (run1), YerevaNN (run3)
- zh2en (8 teams): ai_not_intellegent (run1), Alibuba (run1), baidu_translation (run1), Huawei United (run3), OOM_20 (run1), TMT (run1), TRAMECAT (run1), Wei-Bot (run1)

In addition to the above teams, we also considered the reference translation in the manual validation. We refer to these translations as validation *items* from here on. The selected sentences and abstracts were uploaded into the Appraise tool (Federmann, 2010) for manual validation. The valida-

Teams	Runs	de2en	es2en	fr2en	it2en	pt2en	ru2en	zh2en
Alibuba	Run1	-	-	-	-	-	-	0.2425*
Huawei_United	Run1	0.3897	-	0.4445	0.4974	-	0.4303	0.3378
	Run2	0.4146	-	0.4445*	0.4974*	-	0.4303*	0.3397
	Run3	0.4133	-	-	-	-	-	0.3528
Ixamed	Run1	-	0.4072*	-	-	-	-	-
	Run2	-	0.4073	-	-	-	-	-
	Run3	-	0.3999	-	-	-	-	-
NLE	Run1	0.4043	0.5075*	0.4349*	0.5011*	-	-	-
	Run2	0.4059	-	-	-	-	-	-
	Run3	0.4094*	-	-	-	-	-	-
OOM_20	Run1	-	-	-	-	-	-	0.3483*
	Run2	-	-	-	-	-	-	0.3473*
Sheffield	Run1	-	0.4624*	0.3514*	0.2276*	0.5334*	0.2936*	-
TMT	Run1	0.4165*	-	-	-	-	-	0.3048*
	Run2	0.4037	-	-	-	-	-	0.2893
	Run3	0.4080	-	-	-	-	-	0.2765
TRAMECAT	Run1	-	0.4468*	0.3477*	-	-	0.3707*	0.1688*
TXT	Run1	-	-	-	-	-	-	0.3048*
	Run2	-	-	-	-	-	-	0.2893
	Run3	-	-	-	-	-	-	0.2765
UNICAMP_DL	Run1	-	-	-	-	0.4988*	-	-
	Run2	-	-	-	-	0.4361	-	-
UNICAM	Run1	0.3962	0.4662	-	-	-	-	-
	Run2	0.3979	0.4640	-	-	-	-	-
	Run3	0.3963*	0.4657*	-	-	-	-	-
Wei-Bot	Run1	-	-	-	-	-	-	0.4009*§
	Run2	-	-	-	-	-	-	0.3946§
YerevaNN	Run1	0.4129	-	-	-	-	-	-
	Run2	0.4144	-	-	-	-	0.4331	-
	Run3	0.4128*	-	-	-	-	0.4321*	-
ai_not_intellegent	Run1	-	-	-	-	-	-	0.3357
	Run2	-	-	-	-	-	-	0.3226
	Run3	-	-	-	-	-	-	0.3323
baidu_translation	Run1	-	-	-	-	-	-	0.2494
nrpu-fjwu	Run1	-	-	0.2624*	-	-	-	-
	Run2	-	-	0.2273	-	-	-	-
	Run3	-	-	0.2041	-	-	-	-
one_connect_000	Run1	-	-	-	-	-	-	0.2238*
Baseline	-	0.3470	0.3534	0.3458	0.4588	0.4549	0.2984	0.1561

Table 11: BLEU scores for "OK" aligned test sentences, into English. * Indicates the primary run as indicated by the participants. § Runs submitted after the official test period.

	de2en	en2de	ru2en	en2ru	zh2en	en2zh
AFRL	-	0.2652	0.2895	-	-	-
ariel197197	-	-	0.2999	0.2270	-	-
DeepMind	-	-	-	-	0.2907	-
DiDi_NLP	-	-	-	-	-	-
eTranslation	-	0.257	0.3077	-	-	-
Huoshan_Translate	0.3287	0.2781	-	-	-	-
Online-A	0.3164	0.2649	0.2926	0.2115	0.2413	0.3431
Online-B	0.3342	0.2851	0.3514	0.2594	0.3041	0.3817
Online-G	0.3402	0.2536	0.335	0.2934	0.2854	0.3587
Online-Z	0.2786	0.2172	0.2379	0.1903	0.2162	0.2867
OPPO	0.3287	0.2792	0.3241	0.2566	0.3012	0.3908
PROMT_NMT	0.3100	0.2648	0.3230	0.2502	-	-
SJTU-NICT	-	-	-	-	0.3034	0.4159
Tencent_Translation	-	-	-	-	-	-
Tohoku-AIP-NTT	0.3411	0.2797	-	-	-	-
UEDIN	0.3160	0.2411	-	-	-	-
WMTBiomedBaseline	0.2865	0.2443	-	-	0.1529	-
yolo	0.0022	-	-	-	-	-
zlabs-nlp	0.2516	0.2225	0.2403	0.2016	0.2159	0.2868
Total	12	13	10	8	9	7

Table 12: BLEU scores for news test sentences

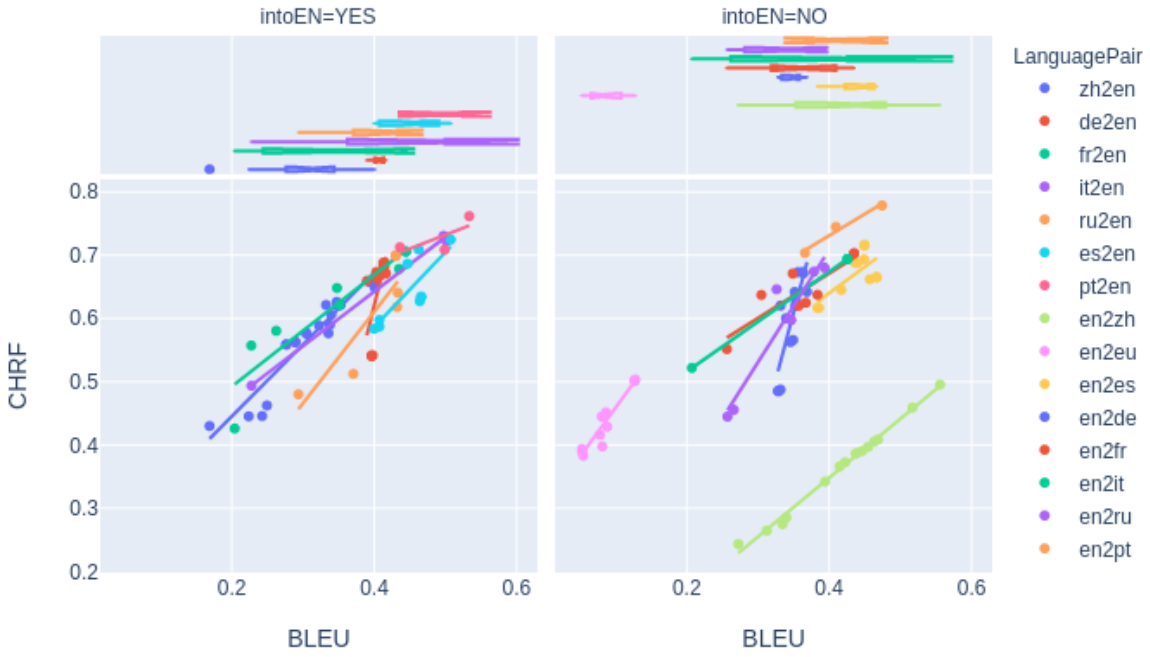

Figure 1: Fitted plot of BLEU vs. chrF scores for “OK" aligned test sentences, into English (left) and from English (right). The top section of the figure shows box plots of the BLEU score distribution for each language pair.

	de2en	en2de	ru2en	en2ru	zh2en	en2zh
AFRL	-	0.3193	0.3847	-	-	-
ariel197197	-	-	0.3911	0.3075	-	-
DeepMind	-	-	-	-	0.3015	-
DiDi_NLP	-	-	-	-	-	-
eTranslation	-	0.3097	0.4008	-	-	-
Huoshan_Translate	0.3915	0.3401		-	-	-
Online-A	0.3739	0.3229	0.3799	0.2926	0.2515	0.3723
Online-B	0.4009	0.3471	0.4711	0.3611	0.3210	0.4138
Online-G	0.3994	0.3086	0.4410	0.4089	0.2906	0.3897
Online-Z	0.3347	0.2546	0.3154	0.2587	0.2203	0.3096
OPPO	0.3915	0.3378	0.4239	0.3529	0.3166	0.4227
PROMT_NMT	0.3693	0.3167	0.4199	0.3434	-	-
SJTU-NICT	-	-	-	-	0.3217	0.4508
Tencent_Translation	-	-	-	-	-	-
Tohoku-AIP-NTT	0.4016	0.3388	-	-	-	-
UEDIN	0.3727	0.2922	-	-	-	-
WMTBiomedBaseline	0.3727	0.2864	-	-	0.1565	-
yolo	0.0026	-	-	-	-	-
zlabs-nlp	0.2961	0.2711	0.3188	0.2815	0.2277	0.3035
Total	12	13	10	8	9	7

Table 13: BLEU scores for news test “OK” sentences

tors were native speakers of the target language and had good knowledge of the source language. Each validator was presented with the source sentence (or abstract), and two candidate translations, either from two teams or from one team and the reference translation. The goal of the validator was to decide whether one translation was better than the other or whether they were of similar quality. Sentences could be skipped if the translations seemed to refer to different source sentences. Results for the manual validation are presented in various tables as summarized below:

- en2de and de2en: Table 16
- en2es and es2en: Table 17

Teams	Runs	BLEU
DCU-MT	Run1	0.0867
	Run2	0.0825
	Run3	0.0808*
Elhuyar_NLP	Run1	0.1271*
	Run2	0.1279
	Run3	0.1268
Ixamed	Run1	0.0815*
	Run2	0.0782
	Run3	0.0884
UTS_NLP	Run1	0.0530*
	Run2	0.0549
	Run3	0.0528
Baseline	-	0.0596

Table 14: Results for the abstract test set (en2eu). * indicates the primary run as indicated by the participants.

Teams	Runs	Accuracy	BLEU
Elhuyar_NLP	run1*	0.78	0.7373
	run2	0.77	0.7356
	run3	0.75	0.7229
ADAPT	run1	0.73	0.7083
	run2	0.76	0.7239
	run3	0.75	0.7179
UTS_NLP	run1*	0.73	0.7115
	run2	0.73	0.7122
	run3	0.73	0.7085
Ixamed	run1	0.12	0.1314
	run2*	0.08	0.0721
	run3	0.13	0.1481

Table 15: Results for the terminology test set (en2eu). * indicates the primary run as indicated by the participants.

- en2fr and fr2en: Table 18
- en2it and it2en: Table 19
- en2pt and pt2en: Table 20
- en2ru and ru2en: Table 21
- en2zh and zh2en: Table 22

We identified the item of each pairwise comparison (if any) that performed better (cf. respective tables) and ran a Wilcoxon Signed-Rank Test using the Python `scipy` library (Virtanen et al., 2020). We consider all comparisons for two particular items over all validated segments (abstracts and sentences), except for skipped segments. The test was calculated for the abstracts and the sentences and we mark in bold in the respective tables if any of them was found to be significant, (p-value< 0.05), otherwise, the two items were considered to be similar. For the language pairs validated by two experts (i.e., es2en and pt2en), we only consider one item of the pairwise comparison to be superior to the other when at least two of the four comparisons (2x for the abstracts, 2x for the sentences) were statistically significant.

To rank the systems, we assign points to each item: 3 points if superior to the opponent, 1 point when they are similar and no points if inferior to the opponent. Based on this methodology, we ranked the systems and the reference translations as summarized below (the obtained points are shown in parentheses):

- en2de: UNICAM (1) < reference (5) < YerevaNN (6) < Huawei-United (7) = NLE (7) < TMT (9)
- en2es: reference (2) < TRAMECAT (4) < Sheffield (5) < Ixamed (6) = UNICAM (6) < Elhuyar_NLP (11)
- en2fr: Sheffield (2) < TRAMECAT (3) = LIMSI (3) < nrpu-fjwu (5) < Huawei United (12) < reference (15)
- en2it: Sheffield (0) < reference (4) = Huawei United (4)
- en2pt: UNICAMP_DL (0) < reference (3) = Sheffield (3)
- en2ru: Sheffield (1) < TRAMECAT (2) < Huawei United (4) < YerevaNN (9) < reference (12)
- en2zh: TRAMECAT (1) < TMT (6) < baidu (10), ai_not_intellegent (10) < Wei-Bot (12) = OOM (12) = Huawei United (12) = Alibuba (12) = reference (12)
- de2en: UNICAM (2) < TMT (5) = reference (5) < Huawei United (7) = YerevaNN (7) = NLE (7)
- es2en: reference (5) = Ixamed (5) = NLE (5) = Sheffield (5) = TRAMECAT (5) = UNICAM (5)
- fr2en: npru-fjwu (0) < TRAMECAT (4) = Sheffield (4) < reference (11) = NLE (11) = Huawei United (11)
- it2en: Sheffield (0) < reference (4) < NLE (5) < Huawei United (7)

- pt2en: reference (2) = UNICAMP_DL (2) = Sheffield (2)
- ru2en: Sheffield (0) < TRAMECAT (3) < reference (8) = YerevaNN (8) = Huawei United (8)
- zh2en: TRAMECAT (0) < TMT (6) < Alibuba (8) < ai_not_intellegent (10) = OOM (10) = reference (10) < baidu (14) = Wei-Bot (14) = Huawei United (14)

The performance of the reference translations varied from being inferior to all runs that were validated, to being superior to all of them. However, for many language pairs, it was as good as the best runs. We summarize the performance of the reference translation below:

- Inferior to all submissions: en2es
- Superior to one or more submissions: en2de, it2en, zh2en, de2en
- Similar to the best submissions: en2it, en2pt, en2zh, pt2en, fr2en, es2en, ru2en
- Superior to all submissions: en2fr, en2ru

In general, the runs that obtained the best scores in the automatic evaluation were also the ones better ranked in the manual evaluation. We highlight the interesting differences to the automatic evaluation below:

en2es: Even though the UNICAM run obtained a slightly higher BLEU score than the ElhuyarNLP one, the latter was ranked much higher. Further, the Ixamed run was ranked reasonably high, even though it obtained the lowest BLEU score.

en2fr: The nrpu-fjwu run was ranked higher the LIMSI run, even though its BLEU score was slightly lower than the one from LIMSI.

en2zh: The run from Alibuba was ranked together with the highest runs, even though its BLEU score was the second lowest one. The Wei-Bot runs was considered as good as some other ones, even though its BLEU score was considerably higher.

de2en: While we did not observe a large difference in the BLEU scores for the runs, three teams (Huawei United, YerevaNN, NLE) were ranked higher than the other two (UNICAM and TMT).

pt2en: While the Sheffield run obtained a higher BLEU score, runs from the Sheffield and UNICAMP_DL were ranked as similar. However, as stated above, we could not perform a manual validation over a larger set of abstracts.

zh2en: The same differences that we observed for en2zh also occurred for zh2en.

es2en: Even though our evaluation relied on very few abstracts, the results confirmed the ones obtained in the automatic evaluation: all systems seem indeed to have a similar quality.

6.2 Basque abstracts

For the human evaluation of the systems that participated in the English-Basque scientific translation, we only carried out the evaluation at sentence-level. We randomly sampled a total of 100 sentences. The runs that we considered from each team are:

- en2eu (4 teams): DCU-MT (run1), Elhuyar_NLP_team (run2), Ixamed (run3), UTS_NLP (run2)

The results of the human evaluation carried out with Appraise are in Table 23, and like in the MEDLINE evaluation, bold numbers indicate a significance difference between the systems after running a Wilcoxon Signed-Rank test. The final ranking of the systems is as follows:

- en2eu: UTS_NLP (0) < DCU-MT (4) = Ixamed (4) < Elhuyar_NLP_team (10) = reference (10)

Similar to what was observed in the MEDLINE evaluation, ranking of the human evaluation matched the ranking of the automatic evaluation.

7 Discussion

In this section we present insights from the automatic and manual validations. We also reflect on the new processes introduced this year in the workflow of the task.

7.1 Analysis of results and methods

Systems submitted to the biomedical task. Figure 1 shows the correlation between BLEU and chrF scores. The use of the survey was helpful to collect specific features of the systems in order to compare the methods used. However, the variety of resources leveraged by the different teams as well as the variety of information reported about

Language	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
en2de	TMT-Huawei	10	3	1	1	104	22	48	26
	TMT-YerevaNN		**4**	1	0		22	53	26
	TMT-NLE		3	1	1		24	62	15
	TMT-UNICAM		**4**	1	0		30	54	17
	TMT-reference		3	0	2		26	45	29
	Huawei-YerevaNN		0	4	1		13	78	11
	Huawei-NLE		2	1	2		21	57	24
	Huawei-UNICAM		1	3	1		**35**	55	12
	Huawei-reference		1	2	2		20	56	26
	YerevaNN-NLE		1	2	2		22	62	19
	YerevaNN-UNICAM		2	1	2		**29**	59	15
	YerevaNN-reference		3	0	2		21	59	23
	NLE-UNICAM		**4**	1	0		24	66	13
	NLE-reference		2	1	2		18	60	25
	UNICAM-reference		2	0	3		18	56	29
de2en	Huawei-YerevaNN	7	1	2	4	50	9	25	16
	Huawei-reference		3	1	3		13	24	13
	Huawei-UNICAM		**4**	3	0		**17**	26	7
	Huawei-TMT		2	1	4		14	27	9
	Huawei-NLE		3	3	1		10	33	7
	YerevaNN-reference		5	1	1		18	21	11
	YerevaNN-UNICAM		5	1	1		**20**	23	7
	YerevaNN-TMT		2	3	2		11	33	6
	YerevaNN-NLE		2	2	3		11	31	8
	reference-UNICAM		4	2	1		19	20	11
	reference-TMT		4	0	3		15	20	15
	reference-NLE		3	0	4		13	26	11
	UNICAM-TMT		1	1	5		8	28	14
	UNICAM-**NLE**		1	2	4		1	36	**13**
	TMT-NLE		2	1	4		5	36	9

Table 16: Manual validation for the en2de and de2en of the MEDLINE abstracts test set. The sum of the values for the sentences does not sum up to the expected value for some rows because some sentences might have been skipped. 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 resources (see Table 7, 8 and 9) make it difficult to directly compare resource use in terms of type or even size. For example, some teams reported the size of their parallel datasets in terms of GB of text, some the number of aligned sentences and sometimes they provided an overall size of resources used for several language pairs.

Biomedical datasets as test suites in the news task. Overall, the best performance on the biomedical datasets was obtained by systems submitted to the biomedical task. These results suggest that domain-specific systems can offer a substantial increase in BLEU score when translating biomedical text. The performance offered by some of the news systems (e.g., Online-B, Online G) was quite high, but it has to be noted that we do not know what training data those system used, and there is no guarantee that our test sentences were not included.

We can also note that whereas no team participated both in the news and biomedical task, we submitted some of our baselines to the news task under the name *WMTbiomedBaseline*. Interestingly, our de2en baseline performed much better there (+2.5 BLEU) on the same text. This is due to supplementary processing: each paragraph to be translated was split into sentences, the sentences were translated one by one, then the results where joined back into a single paragraph. This was not done for the baseline submission to our biomedical translation task, under the assumption that the texts to translate are single-sentence (now invalidated). For the multi-sentence paragraphs, our baselines (as sent to the biomedical task) sometimes contained only the translation of the first sentence, thus leading to a decrease in BLEU score.

7.2 New additions to the workflow of the task

This year, we introduced a number of new processes into the task workflow. First, we performed manual validation of the sentence alignment for three language pairs. This resulted in higher quality alignment, and should be continued. Second, we attempted to split the test sets for the en/fr language

Language	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
en2es	TRAMECAT-UNICAM	9	0	7	2	104	8	82	13
	TRAMECAT-Ixamed		1	7	1		8	85	10
	TRAMECAT-reference		4	2	2		13	81	10
	TRAMECAT-**ElhuyarNLP**		1	7	1		1	93	**10**
	TRAMECAT-Sheffield		3	6	0		5	90	8
	UNICAM-Ixamed		4	4	1		8	90	6
	UNICAM-reference		2	5	2		**13**	87	4
	UNICAM-ElhuyarNLP		2	7	0		4	94	6
	UNICAM-Sheffield		1	6	2		4	93	7
	Ixamed-reference		**5**	4	0		8	83	13
	Ixamed-**ElhuyarNLP**		0	5	**4**		4	93	7
	Ixamed-Sheffield		0	7	2		4	94	6
	reference-**ElhuyarNLP**		0	4	**5**		6	89	9
	reference-Sheffield		1	4	4		6	88	10
	ElhuyarNLP-Sheffield		2	6	1		4	98	2
es2en	Sheffield-Ixamed	2	2/0	0/2	0/0	14	**9**/6	5/6	0/2
	Sheffield-TRAMECAT		1/1	0/1	1/0		6/4	4/8	4/2
	Sheffield-NLE		0/9	0/2	2/0		2/1	7/11	5/2
	Sheffield-UNICAM		1/9	0/2	1/0		4/3	7/10	3/1
	Sheffield-reference		0/1	0/1	2/0		0/0	4/12	**10**/2
	Ixamed-TRAMECAT		1/1	0/0	1/1		1/2	4/8	**9**/4
	Ixamed-NLE		0/1	0/0	2/1		0/1	3/7	**11**/6
	Ixamed-UNICAM		0/0	0/0	2/2		1/0	7/7	6/**7**
	Ixamed-reference		0/1	0/0	2/1		1/2	1/7	**12**/5
	TRAMECAT-NLE		1/0	0/2	1/0		1/0	8/12	5/2
	TRAMECAT-UNICAM		0/0	0/1	2/1		3/1	5/12	6/1
	TRAMECAT-reference		0/1	0/1	2/0		0/0	5/12	**9**/2
	NLE-UNICAM		2/0	0/2	0/0		5/0	8/14	1/0
	NLE-reference		1/1	0/1	1/0		2/0	6/13	6/1
	UNICAM-reference		0/1	0/1	2/0		1/0	6/12	**7**/0

Table 17: Manual validation for the en2es and es2en of the MEDLINE abstracts test set. The sum of the values for the sentences (or abstracts) does not sum up to the expected value for some rows because some sentences (or abstracts) might have been skipped. The better performing system (or reference translation) in each pairwise comparison is depicted in bold, as well as the respective value that has been identified as superior. For es2en, two values are shown: on the left is the validation performed by the English native speaker, and on the right the one from the Spanish native speaker.

pair according to the source language as inferred from MEDLINE metadata. Our experience so far is inconclusive and shows that the initial selection of separate test sets based on source language should be done upstream in the process, as most of the test documents selected had English as the original language. The collection of system information through a survey was effective to collect general comparable information about the systems, especially as the task is growing in number of participants and language pairs offered. However, direct comparison of methods or resources is not necessarily facilitated as authors report information in different ways. A better method for yielding actionable comparisons could be to host a “constrained track" where participants would be requested to use a choice of resources provided in the track.

7.2.1 MEDLINE test sets

We previously presented (cf. Table 2) the results of the manual validation of the automatic alignment that was carried out for the test sets. Here we discuss some of the problems that we found in the automatic alignment for each of the languages.

For all the language pairs, many of the mistakes that we found referred to the titles of the articles, which are usually only available in one of the languages in MEDLINE. Therefore, many of them were correctly aligned to nothing, later identified by the evaluators as being a “NO_ALIGNMENT". However, in some cases, they was incorrectly aligned to the first sentence of the other language, which resulted in them being classified as an “OVERLAP".

The sub-sections which are present in many abstracts, such as “Background" or “Methods" were a cause for trouble. Given their simplicity, they were often correctly aligned. However, in some cases they were aligned to nothing at all (“NO_ALIGNMENT"). In other cases, they were joined to the following or previous sen-

Language	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
en2fr	nrpu-fjwu-**Huawei**	6	0	0	**6**	100	17	19	**64**
	nrpu-fjwu-LIMSI		5	0	1		36	28	36
	nrpu-fjwu-**reference**		0	0	**6**		15	14	**71**
	nrpu-fjwu-TRAMECAT		2	2	2		38	29	33
	nrpu-fjwu-Sheffield		**4**	2	0		41	19	39
	Huawei-LIMSI		**6**	0	0		**61**	23	16
	Huawei-**reference**		1	0	5		16	8	**56**
	Huawei-TRAMECAT		**6**	0	0		**59**	33	8
	Huawei-Sheffield		**6**	0	0		**57**	32	10
	LIMSI-**reference**		0	0	**6**		9	14	**77**
	LIMSI-TRAMECAT		1	3	2		31	32	37
	LIMSI-Sheffield		1	3	2		29	27	43
	reference-TRAMECAT		**6**	0	0		**69**	25	6
	reference-Sheffield		**6**	0	0		**69**	21	8
	TRAMECAT-Sheffield		2	1	3		28	32	38
fr2en	reference-NLE	11	5	2	3	109	36	28	44
	reference-Huawei		3	1	7		37	27	43
	reference-TRAMECAT		8	1	2		**66**	26	15
	reference-Sheffield		8	0	3		**64**	20	23
	reference-nrpu-fjwu		**10**	1	0		**79**	20	8
	NLE-Huawei		5	1	5		28	57	24
	NLE-TRAMECAT		**9**	2	0		**73**	21	15
	NLE-Sheffield		**9**	1	1		**69**	29	11
	NLE-nrpu-fjwu		**11**	0	0		**89**	14	6
	Huawei-TRAMECAT		**11**	0	0		**78**	24	7
	Huawei-Sheffield		**11**	0	0		**70**	30	9
	Huawei-nrpu-fjwu		**11**	0	0		**87**	19	3
	TRAMECAT-Sheffield		4	2	5		37	29	43
	TRAMECAT-nrpu-fjwu		**8**	2	1		**64**	28	17
	Sheffield-nrpu-fjwu		8	0	3		**65**	21	23

Table 18: Manual validation for the en2fr and fr2en of the MEDLINE abstracts test set. The sum of the values for the sentences does not sum up to 109 for some rows because some sentences might have been skipped. The better performing system (or reference translation) in each pairwise comparison is depicted 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-reference	11	5	3	3	100	32	45	23
	huawei-sheffield		**10**	0	0		**80**	14	0
	reference-sheffield		**9**	0	0		**80**	12	4
it2en	sheffield-**reference**	9	0	0	**9**	100	5	1	**94**
	huawei-reference		6	1	2		**46**	37	17
	nle-reference		6	2	1		40	35	25
	huawei-sheffield		**9**	0	0		**98**	2	0
	sheffield-**nle**		0	0	**9**		1	3	**96**
	huawei-nle		3	4	2		27	53	20

Table 19: Manual validation for the en2it and it2en of the MEDLINE abstracts test set. The sum of the values for the sentences (or abstracts) does not sum up to the expected value for some rows because some sentences (or abstracts) have been skipped. 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.

tence and aligned to a sentence in the other language, which did not contain the corresponding sub-section. Such cases were classified as either "SOURCE_GREATER_TARGET", or "TARGET_GREATER_SOURCE".

Comparing one sentence in one language that was automatic aligned to two or more sentences also sometimes caused mistakes. While most of the information is present in both languages, there were always differences between them, and more information in the language for which the alignment tool joined more than one sentence. Depending on the case, the alignment was classified as either "SOURCE_GREATER_TARGET", or "TARGET_GREATER_SOURCE".

Finally some alignments were classified as being "SOURCE_GREATER_TARGET", "TARGET_GREATER_SOURCE", or "OVERLAP"

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
en2pt	**reference**-UNICAMP_DL	13	**9**	3	1	107	**37**	56	14
	reference-Sheffield		6	3	4		18	69	20
	UNICAMP_DL-**Sheffield**		0	2	**11**		8	63	**36**
pt2en	reference-UNICAMP_DL	4	4/2	0/2	0/0	47	18/7	18/35	11/5
	reference-Sheffield		1/1	1/2	2/1		10/2	28/37	9/8
	UNICAMP_DL-Sheffield		0/0	1/1	3/3		6/**38**	29/9	12/0

Table 20: Manual validation for the en2pt and pt2en of the 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. For pt2en, two values are shown: on the left is the validation performed by the English native speaker, and on the right the one from the Portuguese native speaker.

Language	Pair	Abstracts				Sentences			
		Total	**A>B**	**A=B**	**A<B**	**Total**	**A>B**	**A=B**	**A<B**
en2ru	Huawei-**YerevaNN**	6	1	1	5	66	5	41	**18**
	Huawei-Sheffield		**5**	2	0		**29**	24	12
	Huawei-**reference**		0	2	**5**		1	40	**24**
	Huawei-TRAMECAT		3	3	1		22	31	11
	YerevaNN-Sheffield		**7**	0	0		**36**	28	1
	YerevaNN-**reference**		0	4	3		6	45	**15**
	YerevaNN-TRAMECAT		4	1	2		**26**	35	5
	Sheffield-**reference**		0	0	**7**		2	29	**35**
	Sheffield-TRAMECAT		0	5	2		10	38	18
	reference-TRAMECAT		**7**	0	0		**35**	30	1
ru2en	**Huawei**-Sheffield	6	**7**	0	0	58	**38**	15	4
	Huawei-reference		1	6	0		7	46	5
	Huawei-TRAMECAT		**6**	1	0		**24**	29	5
	Huawei-YerevaNN		2	5	0		12	40	6
	Sheffield-**reference**		0	0	**7**		1	17	**40**
	Sheffield-**TRAMECAT**		0	5	2		7	31	**20**
	Sheffield-**YerevaNN**		0	1	**6**		5	17	**34**
	reference-TRAMECAT		**4**	3	0		**19**	34	5
	reference-YerevaNN		2	4	1		9	39	10
	TRAMECAT-**YerevaNN**		0	2	**5**		8	31	18

Table 21: Manual validation for the en2ru and ru2en of the MEDLINE abstracts test set. The sum of the values for the sentences does not sum up to the expected value for some rows because some sentences might have been skipped. 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.

when small details were present in just one of the languages. For instance, one example lacked the information about the p-value, i.e., "($p < 0.05$)", for one of the languages. For another abstract, the sentence in one language referred to the expression "in the city", while the one in the other language explicitly included the name of the city, i.e., "in Paris". It was common that a variety of small details or additional information which were not equally included for both languages.

7.2.2 Basque Abstracts test set

The alignment between the sentences for the abstracts in Basque and English was also carried out manually. Twelve sentences in Basque lack their translation in English and so these sentences in Basque were removed, resulting in the final test set of 375 pairs. The translations produced by the authors of the abstracts are not literal, and in some cases the information given in both languages is different. For example, in two consecutive sentences in an abstract about the listeriosis disease, we have these sentence pairs: First sentence (sentence 1):

- en: In recent years, we have detected a significant increase in the number of cases in Gipuzkoa.
- eu: *Azken urteotan, Gipuzkoan,* ***listeriosiaren*** *intzidentziaren igoera esanguratsua atzeman da.*
  'In recent years, in Gipuzkoa, there has been a significant increase in the incidence of listeriosis'

Following sentence (sentence 2):

- en: **Listeriosis** is uncommon in the general population, but it is far more frequent in pregnant women and newborns.

Pair	en2zh - Abstracts			Pair	zh2en - Abstracts		
	A>B	A=B	A<B		A>B	A=B	A<B
reference-TRAMECAT	**16**	1	0	**reference**-TRAMECAT	**19**	1	0
reference-baidu	9	2	4	reference-baidu	6	2	6
reference-TMT	**16**	0	1	reference-TMT	13	2	5
reference-Wei-Bot*	9	1	7	reference-Wei-Bot*	5	4	11
reference-ai_not_intellegent	9	1	7	reference-ai_not_intellegent	10	0	10
reference-OOM	8	2	7	reference-OOM	5	3	12
reference-Huawei	10	1	6	reference-Huawei	4	6	10
reference-Alibuba	7	1	6	reference-Alibuba	7	1	6
TRAMECAT-**baidu**	0	0	**14**	TRAMECAT-**baidu**	0	1	**13**
TRAMECAT-TMT	4	2	11	TRAMECAT-**TMT**	1	1	**18**
TRAMECAT-**Wei-Bot***	0	1	**16**	TRAMECAT-**Wei-Bot***	1	0	**19**
TRAMECAT-**ai_not_intellegent**	0	0	**17**	TRAMECAT-**ai_not_intellegent**	0	2	**18**
TRAMECAT-**OOM**	0	0	**17**	TRAMECAT-**OOM**	0	0	**20**
TRAMECAT-**Huawei**	0	0	**17**	TRAMECAT-**Huawei**	0	0	**20**
TRAMECAT-**Alibuba**	0	0	**14**	TRAMECAT-**Alibuba**	0	0	**14**
baidu-TMT	9	2	4	baidu-TMT	8	1	5
baidu-Wei-Bot*	0	14	1	baidu-Wei-Bot*	0	14	0
baidu-ai_not_intellegent	5	5	5	**baidu**-ai_not_intellegent	**9**	3	2
baidu-OOM	0	13	2	baidu-OOM	0	13	2
baidu-Huawei	3	7	5	baidu-Huawei	4	8	2
baidu-Alibuba	6	6	2	**baidu**-Alibuba	**9**	3	2
TMT-**Wei-Bot***	3	2	**12**	TMT-**Wei-Bot***	3	2	**15**
TMT-ai_not_intellegent	1	3	**13**	TMT-**ai_not_intellegent**	1	6	**13**
TMT-**OOM**	2	1	**14**	TMT-**OOM**	3	3	**14**
TMT-**Huawei**	2	1	**14**	TMT-**Huawei**	3	2	**15**
TMT-**Alibuba**	2	2	**11**	TMT-Alibuba	3	1	10
Wei-Bot*-ai_not_intellegent	6	7	4	**Wei-Bot***-ai_not_intellegent	**12**	6	2
Wei-Bot*-OOM	0	16	1	Wei-Bot*-OOM	0	18	2
Wei-Bot*-Huawei	6	7	4	Wei-Bot*-Huawei	7	7	6
Wei-Bot*-Alibuba	5	7	3	Wei-Bot*-Alibuba	3	3	8
ai_not_intellegent-OOM	2	7	8	ai_not_intellegent-OOM	2	5	13
ai_not_intellegent-Huawei	3	8	6	ai_not_intellegent-Huawei	4	6	10
ai_not_intellegent-Alibuba	1	13	1	ai_not_intellegent-Alibuba	0	14	0
OOM-Huawei	8	7	2	OOM-Huawei	6	11	3
OOM-Alibuba	6	6	3	OOM-Alibuba	10	1	3
Huawei-Alibuba	6	6	3	**Huawei**-Alibuba	**9**	4	1

Table 22: Manual validation for the en2zh and zh2en of the MEDLINE abstracts test set. The evaluation was carried out only for abstracts: 17 for en2zh, and 20 for zh2en. The sum of the values for the abstracts does not sum up to the expected value for some rows because some abstracts might have been skipped. The better performing system (or reference translation) in each pairwise comparison is shown in bold, as well as the number of times this system was superior. The system identified with an * cannot be fully compared to the other systems.

Pair	Sentences			
	Total	A>B	A=B	A<B
reference-UTS_NLP	100	**91**	7	2
reference-Ixamed	100	**68**	13	19
reference-Elhuyar_NLP	100	37	33	30
reference-DCU-MT	100	**75**	10	15
Ixamed-UTS_NLP	100	**60**	11	29
Ixamed-**Elhuyar_NLP**	100	17	25	**58**
Ixamed-DCU-MT	100	51	7	42
Elhuyar_NLP-UTS_NLP	100	**94**	6	0
Elhuyar_NLP-DCU-MT	100	**67**	24	9
DCU-MT-UTS_NLP	100	**74**	17	9

Table 23: Manual validation of the en2eu 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.

- eu: *Arrisku- taldeen artean, haurdun dauden emakumeak aurkitzen dira.*
  'Risk groups include pregnant women'

In the first sentence pair, the name of the disease is given in Basque, while in the second pair, the mention is given in English. In the second pair, the sentence in English gives more information than the one in Basque. This fact could well affect the

automatic evaluation.

7.3 Quality of the system translations

We discuss below some of the mistakes that we found during the manual validation of the selected runs and the reference translations.

7.3.1 MEDLINE test sets

en (from de) The quality of the translations has substantially improved since last year, with many instances requiring lengthy manual scrutiny to detect slight nuances in the meaning of the translated texts. In some cases, the subject matter of the abstracts presented a real challenge for the manual validator, as some of the translations required deeper background knowledge of medical procedures and terms to evaluate whether or not the translations from the source language were indeed correct. Examples include: (1) the German term *Hyperandrogenämie* was correctly translated to“hyperandrogenemia" (referring to elevated levels of androgen in the blood) or incorrectly to “hyperandrogenism" (refers to the state characterized by elevated levels of androgens); (2) in the context of liver cirrohsis, the “Child-Pugh-Score" was used as a pro-form term for liver cirrohsis disease severity. In this particular case, the correct translation was not even evident until the abstract was evaluated as a whole, since the manual validation of single sentences did not even contain the term *Child-Pugh-Stadium* in the source German sentence; (3) in an ophthalmology abstract, the German phrase *Aufgrund des ausgeprägten Hornhautödems* was correctly and literally translated in one instance as “Due to the pronounced corneal edema" but slightly differently in the other instance as “Due to the pronounced corneal endothelial epithelial decompensation", which may be partially correct in that corneal edema is a clinical feature of corneal endothelial epithelial decompensation. Such an interpretation would be best evaluated by an opthalmologist.

Abbreviations continue to present difficulties for correct translation. For example, in German, *Cephalosporine der 3. Generation* was never correctly translated to “third generation cephalosporins". Also the disease abbreviation *HEED (Hornhaut-Endothel-Epithel-Dekompensation)* could not be translated into English, though the disease was correctly translated in English to “corneal endothelial epithelial decompensation". The abbreviation for *polyzystische Ovarsyndrom (PCOS)* was incorrectly interpreted as a plural (“PCOs") in one translation.

Some specific medical terms were literally translated from the German source words, but resulted in an unusual or rare choice of words in English. For example, *Darm-Hirn-Achse* literally translated to “bowel-brain axis" instead of “brain-gut axis", *Adipositas* directly to “adiposity" vs. “obesity", *Mikrobiomtransfers* to “microbiome transfer" vs. “microbiota transplantation", *Kupfer-Instrauterinpessar* to “IUP" instead of “intrauterine device (IUD)". In these examples, the translations are in principle still understandable, yet awkward in English.

In some cases, choosing an English synonym of a translated German word altered the original German meaning entirely. For example, the German phrase *abgeschlossenen und laufenden kontrollierten Studien* was translated into “terminated and ongoing controlled trials" as well as “completed and ongoing controlled studies", whereby the use of the adjective “terminated" in this specific context implies that the clinical trial was prematurely stopped, possibly due to ethical, financial, safety or efficacy concerns. In this context, “completed" is the better adjective, as it implies that a study protocol was carried out to its scheduled endpoints. Similarly, in the context of raising children, the German *Erziehungserfahrungen* was sometimes translated to “educational experience", rather than the correct term “parenting experiences".

es (from en) This year, five different MT systems competed against the human reference translation for the English to Spanish language pair. The overall quality of all five systems was very good this year, when comparing sentences, being equal to the human translation in many instances.

The handling of acronyms still requires improvement for some of the MT systems, as the treatment vary from inconsistent translation, in the case of abstracts, to wrong use of lower case instead of capital letters as in the following example, correct acronym for *Sistema Único de Salud* (*SUS*) versus *Sistema Único de Salud* (*sus*). There were also some instances of literal translation of terms such as the mistranslation of *severe temperature* as *temperatura severa* when a more correct translation would have been *temperatura grave*.

In long sentences, there were also cases of missing information in the MT systems that affected the overall quality of the translations. In the rare cases where there were no clear issues with the

MT output, the human translation was sometimes more readable and more fluent and therefore the preferred choice in terms of quality. As in the following example:

- Original English text: *The objective was to assess parental knowledge, behaviors, and fears in the management of fever in their children.*
- Tramecat Translation: *El objetivo fue evaluar el conocimiento, comportamientos y miedos de los padres en el manejo de la fiebre en sus hijos.*
- Reference translation: *El objetivo fue evaluar los conocimientos, actitudes y temores de los padres ante la fiebre de sus hijos.*

The noun group elements have greater concordance in the reference translation rendering it more readable and fluent than the tramecat MT system. When comparing the reference abstracts to the MT abstracts, the human translation had higher quality due to its consistency and overall textual coherence. Some systems had issues with term translation consistency, non-fluent text (rare) or missing information (also rare). As mentioned, the MT systems performed very well when compared with one another and with the reference translation, to obtain a good level of quality, but in some cases many of the systems would still require human intervention in terms of post-edition to improve them to publishing quality level.

en (from fr) The overall quality of translations was high, with many perfect translations. Most translation issues arose from unknown vocabulary or an inappropriate use of vocabulary in context. This includes (i) the presence of untranslated French words (*We montrons* as a translation of *nous montrons* 'we show'), (ii) the erroneous translation of subword units, resulting in a merging of units (*tharural* instead of *than rural*), (iii) erroneous translation of context-dependent ambiguous terms (*Study of litter* as a translation of *étude de portée* 'scoping study' as a consequence of a poor translation of the ambiguous word *portée* 'scope, litter (of puppies)') and (iv) a strange translation of unseen source words that may nevertheless share initial subword units with the predicted word (*consumptions of cruels* as a translation of *consommation de crudités* 'consumption of raw vegetables'). A further issue noted was the poor translation of the French pronoun *il* 'it/he' into *he* when this refers to the article itself. The correct translation of these pronouns necessitates taking into account preceding context.

en (from it) The quality of the translations was neatly divided between almost-perfect and very poor, and this is reflected in the relative rankings between validations reported in Table 19. Outright errors in the good translations were rare; occasionally, the subject of a subordinate clause was mistaken. Interestingly, some translations proved capable of appropriately using synonyms and correctly rendering the meaning of the source with a slightly less literal and more idiomatic translation.

en (from zh) The quality of the translations is generally good. Some systems produced translations that provided not only correctness but also more typical English word usage beyond a literal translation. As an example, 不同性别、年龄别和身高别儿童青少年血压评价 was translated more literally by one system as *blood pressure evaluation in children and adolescents of different sexes, ages and heights*, but another system was able to produce a more natural translation: *blood pressure evaluation in children and adolescents by gender, age and height*.

The biggest source of errors is by far the translation of biomedical concepts. Presumably because a concept is not available in a reference dictionary in the target language, the translation systems often resorted to a literal interpretation of the source characters, leading to a translation that ranged from comprehensible to completely incorrect. For instance, a correct translation for 美观协调 is *aesthetic coordination* (in the context of teeth and jaw operations), but an actual and rather literal translation was *good and beautiful are in harmony*, which was still comprehensible. In another example, however, a correct translation of 早期移植物功能不全 was *early graft dysfunction*, but an incorrect translation yanked two characters 植物 (meaning "plants") out of the 3-character term 移植物 (meaning "transplant matter") and produced *early removal of plant functions*, which was completely incorrect.

A second problem area is the skipping of source words or even phrases. For biomedical texts, even skipping one critical word can significantly alter the context of the entire text. Take 老年骨质疏松人群 as an example, whose full translation is *elderly osteoporosis population*. Some translations

omitted the word *elderly*, and that changed the context of the corresponding scientific study.

fr (from en) Overall, the quality of the translations ranged from fair to good and was improved over previous editions of the task. Some aspects previously noted as difficult (e.g., co-reference, acronym definitions) were correctly translated by some of the systems at the sentence level. However, the abstract-level evaluation evidenced overall consistency issues. For example, a procedure correctly described as *cholécystectomie laparoscopique conventionnelle (CLC)* in an introductory sentence could be referred to with a different acronym , e.g., *CCC* in sentences appearing later in the same abstract. Other issues noted in previous editions remained, such as repeated portions of text (up to 96 repetitions of a word pair in one evaluated sentence) and untranslated sections, especially in passages containing complex numerical data. Some issues with technical vocabulary also led to incorrect translations. In the comparison of translation issues exhibited by different systems in the same sentence, a preference was given to medical correctness over grammatical correctness. For example, when comparing:

- Translation A: *L'étude en microscopie multiphotonique montre que,* ***comme on le attendait****, l'émiline-1* ***se colocalise*** *avec l'élastine.*

  and

- Translation B: *L'étude de microscopie multiphotonique montre que,* ***comme attendu****, l'Emiline-1* ***permet de colorer*** *avec de l'Éastine.*

where *comme attendu* (B) is grammatically preferable to *comme on le attendait* (A) as a translation of *as expected* and *se colocalise* (A) is semantically preferable to *permet de colorer* (B) as a translation of *colocalizes*, translation A is assessed as superior to translation B even though neither translation is perfect.

it (from en) The quality of the translations was strongly influenced by the systems (unknown at the time of the evaluation). Some of the translations were almost perfect and the best system was also able to use the correct technical terminology for specialized domains, such as philosophy and medicine. Other translation were partially correct, in the sense that they were understandable but with syntactic or lexical inconsistencies. For example, the term "otherness"– meaning "being different" – was incorrectly translated by the term *estraneità* (meaning "unfamiliarity") rather than the Italian equivalent *alterità*, which conveys the same meaning. Another example specific for the medical domain is the translation of the multi-word unit "visceral adhesions" by *adesivo viscerale* ("visceral sticker" as a literal translation) rather than the correct Italian equivalent *aderenze viscerali*. Finally, some other translations presented non-existent Italian words.

en (from pt) The translations have high fidelity to the source texts, but in terms of natural language style and typical word usage, the translations are clearly lacking, especially in longer sentences. There was a small number of critical errors in translating biomedical concepts, rendering the translation incomprehensible. For example, *acidentes ofídicos* was correctly translated as *snakebite* or as a more pedantic version, *snakebite envenomations*, but one incorrect translation *obscene accidents* was too obscure to hint at the original term. Lexical similarity might have been a contributing factor to errors as well. *Ofidismo* (meaning "snakebite") was translated as *ophidism* (meaning "poisoning caused by snake venom"), which was not an exact translation but still highly relevant. However, an incorrect translation *oblivinism* was, to the best of our knowledge, not an English word.

pt (from en) The translations have improved but none of the texts were perfect, since we also found mistakes in the reference translations. One of the most significant improvements, in comparison to previous years, is the lack of untranslated words; only very few of them were observed. However, one of the frequent problems still remains: poor translations of the acronyms, which are often the ones from the English (source) text. Most of the errors were actually in the small details, such as the best choice of words for a particular concept (e.g., *o processo de morte e morte* as a translation of *process of death and dying*), gender or number coordination (e.g., *na encaminhamento dos pacientes*, *programa de formação específico*), or misplacement of commas. Finally, more errors occurred in longer sentences due to their increased complexity than shorter ones, which tended to be correct.

de (from en) The overall quality of translation was high. In various cases the better translations were chosen based on small nuances, such as no capitalization errors, better ordering of words or sentence structure that sounds slightly more natural to a native speaker. Considering the original German abstracts, sentences often appeared to be freely translated, targeting an identical meaning rather than an exact translation. Therefore, in various cases, the automatic translations outperformed the reference translations, which sometimes lacked some information.

Generally the translation of acronyms appears more difficult. In multiple cases, translations used the English acronym instead of the German version, although the underlying term itself has been translated correctly. Finally we observed that some translations favored very technical terms, while others favored rather simple ones, but both correct. In those cases it is difficult to choose the better translation, if the rest of the sentences have the same quality. Generally we believe that using more complicated words does not mean that the translation of a scientific paper is necessarily better.

zh (from en) While the quality of zh2eh translations (discussed above) was already generally good, the quality of en2zh translations was generally even better in comparison.

Where applicable, a very specific term in English can be left untranslated in English in the Chinese text with good effect. Protein names such as *CD34* and long, complicated chemical names with abbreviations are prominent examples. The participating systems employed different strategies here: some repeated only the original English term, some repeated the English term as well as translated it in Chinese, and some translated it in Chinese but appended the English abbreviation.

In terms of language style, some systems produced more natural Chinese word usage than a literal translation. Take *evidence is strongest* as an example. A correct but linguistically clumsy translation was 证据最强, which means exactly "evidence is strongest." But other systems were able to produce more typical wordings such as 证据最有力 (meaning "evidence has most force") or, even better, 证据最为充分 (meaning "evidence is most sufficient").

The translation of biomedical concepts was again the biggest source of error, and again the problematic translations ranged from comprehensible to completely incorrect. For instance, *positive control* in the context of conducting experiments should be correctly translated as 阳性对照, but some system instead produced 积极的控制, which means "positively or enthusiastically take charge." Some translations were outright incorrect, such as when a simple term *fever* was translated as 百日咳, which means "whooping cough."

en (from ru) The English-Russian task was offered for the first time, with four MT systems participating and competing against the reference translation. The quality of translations were generally good, with two systems producing significantly better results. Translations frequently contained synonyms successfully carrying on the meaning of the source sentence. For example, "травматические поражения" is correctly translated as "traumatic lesions" and "traumatic injury". Observed was a range of translations, where some presented a stylistically more elegant solution then the others. For example, the phrase "reduction of pain syndrome" is better expressed as "reduce the level of pain". There was a small number of errors related to incorrect translation of biomedical key terms, resulting in translation being impractical. A mild example of incorrectly translated terminology is "spinal surgeon" instead of "spinal surgery". Skipping over segments of sentence was observed mainly in sentences with challenging tokenization.

ru (from en) The Russian-English task was offered for the first time, with four MT systems participating and competing against the reference translation. The quality of translations were generally good, with two systems producing significantly better results. Abbreviated disease names tended to cause an issue in translation. Sentences containing definition and the first mention of abbreviation contained the correct abbreviation. In subsequent sentences, the abbreviation was getting transliterated. For example, "chronic endometritis (CE)" is translated as "хроническим эндометритом (ХЭ)". However subsequent sentences refer to "CE" as "КЭ" and not as "ХЭ". Rarely observed were instances with the meaning lost in translation. For example, the source sentence "The biological age of sleep apnea patients exceeded the passport age by 41.3% and comorbid patients by 49.6%." was translated as: "Биологический возраст пациентов с апнозом сна превышал пассажиров на 41.3%, а сопутствующих на 49.6%."

7.3.2 Basque abstracts

The BLEU scores for this subtask are given in Table 14. We have to consider that BLEU scores tend to be low when translating into Basque (Jauregi Unanue et al., 2018), and this can be seen in the results. The best performing system in the automatic evaluation was Elhuyar_NLP, with a BLEU score of 0.1279. Ixamed and DCU-MT have similar performance, with UTS_NLP achieving the lowest BLEU score. In spite of the low BLEU scores, the manual evaluation in Table 23 showed that Elhuyar_NLP was competitive against the reference translation, and was preferred to other systems.

During the manual evaluation, the annotators also observed that sometimes the system produced output in Spanish instead of Basque. This was obviously a mistake when using Spanish as a pivot language, but it may have helped the BLEU scores in some cases, due to shared terminology. In the manual annotation, text in Spanish was penalized.

7.3.3 Basque terminology

As explained in Section 2.2.1, the development set and test set were the same, and this caused the results to be higher than in a real setting.[22] The results in Table 15 show that most systems performed with high accuracy and BLEU scores. Elhuyar_NLP was again the highest performer, with Ixamed producing very low scores, perhaps due to an error in their submission. We did not perform manual evaluation for this subtask.

8 Conclusions

We presented the findings of the fifth edition of the WMT biomedical task. This edition addressed three new languages and test sets that included scientific abstracts and terminologies. We explored news ways of improving our tests and carried out (as in previous editions of the task) both an automatic and a manual validation. Results confirmed the improvements of the runs and for some language pairs, suggested that some runs were on a par with or superior to the reference translations.

Acknowledgments

We would like to thank all participants in the challenges, and especially those who supported us for the manual evaluation.

[22] As a reference, one of the participating systems (UTS_NLP) was able to re-run their system over the real test set. The performance drop was 0.08 for accuracy (from 0.73 to 0.65), and 0.05 for BLEU (from 0.71 to 0.66).

References

UFAL medical corpus 1.0. https://ufal.mff.cuni.cz/ufal_medical_corpus. Accessed: 2018-07-24.

Sadaf Abdul Rauf, José Carlos Rosales Núñez, Minh Quang Pham, and François Yvon. 2020. LIMSI @ WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

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.

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.

Ander Corral and Xabier Saralegi. 2020. Elhuyar submission to the Biomedical Translation Task 2020 on terminology and abstracts translation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi Summary Translation Test Data 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Nicolas Griffon, Matthieu Schuers, Gaëtan Keroelhué, Julien Grosjean, and Stéfan J Darmoni. 2017. Littérature scientifique en santé (LiSSa) : une base de données bibliographiques en français [LiSSa, health scientific literature: a French bibliographic database]. *Rev Prat*, 67:134–138.

Karen Hambardzumyan, Hovhannes Tamoyan, and Hrant Khachatrian. 2020. YerevaNN's Systems for WMT20 Biomedical Translation Task: The Effect of Fixing Misaligned Sentence Pairs. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Inigo Jauregi Unanue, Lierni Garmendia Arratibel, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018. English-Basque statistical and neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Inigo Jauregi Unanue and Massimo Piccardi. 2020. Pretrained Language Models and Backtranslation for English-Basque Biomedical Neural Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

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 the 56th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 116–121, Melbourne, Australia.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. 2020. Lite Training Strategies for Portuguese-English and English-Portuguese Translation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Prashant Nayak, Rejwanul Haque, and Andy Way. 2020. The ADAPT's Submissions to the WMT20 Biomedical Translation Task. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Sumbal Naz, Sadaf Abdul Rauf, Noor e Hira, and Sami Ul Haq. 2020. FJWU participation for the WMT20 Biomedical Translation Task. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

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.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

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

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.

Wei Peng, Jianfeng Liu, Minghan Wang, Liangyou Li, Xupeng Meng, Yangm Hao, and Qun Liu. 2020. Huawei's Submissions to the WMT20 Biomedical Translation Task. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Addressing Exposure Bias With Document Minimum Risk Training: Cambridge at the WMT20 Biomedical Translation Task. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A large parallel corpus of full-text scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Felipe Soares, Mark Stevenson, Diego Bartolome, and Anna Zaretskaya. 2020. ParaPat: The multi-million sentences parallel corpus of patents abstracts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3769–3774, Marseille, France. European Language Resources Association.

Felipe Soares and Delton Vaz. 2020. UoS Participation in the WMT20 Translation of Biomedical Abstracts. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Xabier Soto, Olatz Perez-de Viñaspre, Gorka Labaka, , and Maite Oronoz. 2020. Ixamed's submission description for WMT20 Biomedical shared task: benefits and limitations of using terminologies for domain adaptation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Olatz Perez-de Viñaspre and Maite Oronoz. 2015. Snomed ct in a language isolate: an algorithm for a semiautomatic translation. In *BMC medical informatics and decision making*, volume 15, page S5. Springer.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2020. Tencent AI Lab Machine Translation Systems for the WMT20 Biomedical Translation Task. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.