Findings of the WMT 2019 Biomedical Translation Shared Task: Evaluation for MEDLINE Abstracts and Biomedical Terminologies

Rachel Bawden[*]
School of Informatics,
University of Edinburgh,
Scotland

K. Bretonnel Cohen[*]
Biomedical Text Mining Group
University of Colorado
School of Medicine
Aurora, CO, USA

Cristian Grozea[*]
Fraunhofer Institute
FOKUS,
Berlin, Germany

Antonio Jimeno Yepes[*]
IBM Research Australia
Melbourne, Australia

Madeleine Kittner[*]
Knowledge Management
in Bioinformatics
Humboldt-Universität
zu Berlin, Germany

Martin Krallinger[*]
Barcelona Supercomputing
Center, Spain

Nancy Mah[*]
Charité-Universitätsmedizin,
Berlin-Brandenburger Centrum
für Regenerative Therapien (BCRT)
Berlin, Germany

Aurélie Névéol[*]
LIMSI, CNRS,
Université Paris-Saclay
Orsay, France

Mariana Neves[*]
German Centre for the Protection
of Laboratory Animals (Bf3R),
German Federal Institute for
Risk Assessment (BfR),
Berlin, Germany

Felipe Soares[*]
Barcelona Supercomputing
Center, Spain

Amy Siu[*]
Beuth University of
Applied Sciences,
Berlin, Germany

Karin Verspoor[*]
University of Melbourne,
Australia

Maika Vicente Navarro[*]
Maika Spanish Translator
Melbourne, Australia

Abstract

In the fourth edition of the WMT Biomedical Translation task, we considered a total of six languages, namely Chinese (zh), English (en), French (fr), German (de), Portuguese (pt), and Spanish (es). We performed an evaluation of automatic translations for a total of 10 language directions, namely, zh/en, en/zh, fr/en, en/fr, de/en, en/de, pt/en, en/pt, es/en, and en/es. We provided training data based on MEDLINE abstracts for eight of the 10 language pairs and test sets for all of them. In addition to that, we offered a new sub-task for the translation of terms in biomedical terminologies for the en/es language direction. Higher BLEU scores (close to 0.5) were obtained for the es/en, en/es and en/pt test sets, as well as for the terminology sub-task. After manual validation of the primary runs, some submissions were judged to be better than the reference translations, for instance, for de/en, en/es and es/en.

1 Introduction

Machine translation (MT) holds the promise to unlock access to textual content in a wide range of languages. In the biomedical domain, the bulk of the literature is available in English, which provides two interesting applications for machine translation: first, providing patients, scientists and health professionals with access to the literature in their native language and second, assisting scientist and health professionals in writing reports in English, when it is not their primary language. Furthermore, important health information can be found in the free text of electronic health records and social media. As these sources are increasingly available to patients and health professionals, MT can be leveraged to widen access beyond

[*] The author list is alphabetical and does not reflect the respective author contributions. The task was coordinated by Mariana Neves.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 31–55
Florence, Italy, August 1-2, 2019.

language barriers. Other situations in the health care domain, such as emergency response communications, have expressed the need for translation technologies to improve patient-provider communication (Turner et al., 2019). However, the recurring conclusion of practical studies is that progress is still needed. The goal of this shared task is to bring machine translation of biomedical text to a level of performance that can help with these medical challenges.

In recent years, many parallel corpora in the biomedical domain have been made available, which are valuable resources for training and evaluating MT systems. Examples of such corpora include Khresmoi (Dušek et al., 2017), Scielo (Neves et al., 2016), Full-Text Scientific Articles from Scielo (Soares et al., 2018a), MeSpEn (Villegas et al., 2018), thesis and dissertations (Soares et al., 2018b), and clinical trials (Neves, 2017). These corpora cover a variety of language pairs and document types, such as scientific articles, clinical trials, and academic dissertations.

Many previous efforts have addressed MT for the biomedical domain. Interesting previous work includes a comparison of performance in biomedical MT to Google Translate for English, French, German, and Spanish (Wu et al., 2011). Pecina et al. applied MT for the task of multilingual information retrieval in the medical domain (Pecina et al., 2014). They compared various MT systems, including Moses, Google Translate, and Bing Translate. Later, Pecina et al. utilized domain adaptation of statistical MT for English, French and Greek (Pecina et al., 2015). The field of MT has experienced considerable improvements in the performance of systems, and this is also the case for biomedical MT. Our more recent shared tasks show an increasing number of teams that relied on neural machine translation (NMT) to tackle the problem (Jimeno Yepes et al., 2017; Neves et al., 2018).

We found some commonalities in the work above. On the one hand, clinical vocabularies are under development, as well as data sets based on scientific publications. On the other hand, there is little or no work on languages that do not have typical Indo-European morphology, e.g. in the isolating direction (no Chinese), and in the agglutinating direction (no Hungarian, Turkish, Finnish, Estonian). There is also little previous research in MT for electronic health records (EHR).

The translation of technical texts requires considerable specific knowledge, not only about linguistic rules, but also about the subject of the text that is being translated. The advantage of terminology management can be seen in its important role in the process of acquiring, storing and applying linguistic and subject-specific knowledge related to the production of the target text.

Terminologies can also be extremely useful in data mining pipelines, where one might be interested in identifying specific terms or diseases, for example. In addition, terminologies can be used to improve the quality of machine translation and help in the normalization of vocabulary use. Terminological resources in the field of biomedicine and clinic are of crucial importance for the development of natural language processing systems and language technologies in the field of health, among them the semantic network called Unified Medical Language System (UMLS). This resource contains terminological subsets of a wide variety of subject areas and specialties such as health sciences, life sciences and pharmacology.

For instance, at present only 13% of the concepts included in UMLS have entries for Spanish, while the vast majority of concepts have an equivalent in English. Therefore, one of the coverage expansion strategies is based on the translation of terms related to UMLS entries from English into Spanish.

Over the past three years, the aim of the biomedical task at WMT has been to focus the attention of the community on health as a specialized domain for the application of MT (Bojar et al., 2016; Jimeno Yepes et al., 2017; Neves et al., 2018). This forum has provided a unique opportunity to review existing parallel corpora in the biomedical domain and to further develop resources in language pairs such as English and Chinese, French, Spanish, Portuguese (Névéol et al., 2018).

In this edition of the shared task, we continued this effort and we addressed five language pairs in two translation directions, as follows: Chinese/English (zh/en and en/zh), French/English (fr/en and en/fr), German/English (de/en and en/de), Portuguese/English (pt/en and en/pt), and Spanish/English (es/en and en/es). Herein we describe the details of the fourth edition of the WMT Biomedical Task which includes the following:

- construction of training data and the official test sets, including statistics and an evalua-

tion of the quality of the test sets (Section 2);

- a description of the three baselines that we developed for comparison (Section 3);
- an overview of the participating teams and their systems (Section 4);
- the results obtained by the submitted runs based on our automatic evaluation (Section 5);
- the results of the manual evaluation of selected translations from each team (Section 6);
- and a discussion of various topics, especially the quality of the test sets and of the automatic translations submitted by the teams (Section 7).

2 Training and Test Sets

We made training and test sets available to support participants in the development and evaluation of their systems. We provided two types of test set, scientific abstracts from Medline and terms from biomedical terminologies. Both data and test sets are available for download.[1] Table 1 provides some basic characteristics of the training and test sets, and we provide details of their construction in this section.

2.1 Medline training and test sets

We provided training data based on Medline data for eight of the language pairs that we addressed, namely, fr/en, en/fr, de/en, en/de, pt/en, en/pt, es/en, and en/es. We released test sets for all 10 language pairs, which are the official test sets used for the shared task. The creation of the Medline training and test sets was as follows.

Document retrieval. For the training data, we downloaded the Medline database[2] that included the citations available until the end of 2018. For the test sets, we downloaded the Medline update files available for 2019 until the end of February.

XML parsing. We parsed the Medline files using a Python XML library.[3] Based on the metadata available, we selected the citations that contained abstracts both in English and in at least one of the foreign languages addressed in the shared task, namely, Chinese (zh), French (fr), German (de), Portuguese (pt), and Spanish (es).

Language detection. Even though the citations in Medline include the language of the abstracts, we found some mistakes in the data from last year in which some abstracts were tagged with the wrong language, e.g. Italian instead of German. Therefore, we automatically detected the language of the article using the Python langdetect library.[4] For instance, when building the training data, we detected a total of 156 abstracts that were identified with the wrong language. For the training data, this was the data that was released to the participants after removal of the abstracts in the wrong language. When building the test sets, we kept only 100 articles for each language pair, i.e. 50 articles for each direction es/en and en/es.

Sentence splitting. For the test sets, we considered only the abstracts in the Medline citations and segmented them into sentences, which is a necessary step for automatic sentence alignment. For all language pairs except for zh/en, we used the syntok Python library[5]. For zh/en, we used LingPipe's Medline-specific API[6] to segment the English abstracts. Splitting the Chinese ones by the language-specific period punctuation "。" (using our own script) was sufficient.

Sentence alignment. For the test sets in all language pairs except for zh/en, we automatically aligned the sentences using the GMA tool.[7] We relied on the same configuration and stopword lists used for the test sets in 2018 (Neves et al., 2018). For zh/en, we used the Champollion tool[8], also relying on the same configurations and stopword lists used in 2018.

Manual validation. We performed a manual validation of the totality of the aligned sentences in the test sets using the Quality Checking task in

[1] https://drive.google.com/drive/u/0/folders/1x4689LkvdJTyAxsB6tYu12MJzxgiyDZ_

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

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

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

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

[6] http://alias-i.com/lingpipe/demos/tutorial/sentences/read-me.html

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

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

Language pairs	Medline training		Medline test		Terminology test
	Documents	Sentences	Documents	Sentences	Terms
de/en	3,669	40,398	50	589	-
en/de			50	719	-
es/en	8,626	100,257	50	526	-
en/es			50	599	6,624
fr/en	6,540	75,049	50	486	-
en/fr			50	593	-
pt/en	4,185	49,918	50	491	-
en/pt			50	589	-
zh/en	-	-	50	283	-
en/zh			50	351	-

Table 1: Number of documents, sentences, and terms in the training and test sets.

the Appraise tool. We present statistics concerning the quality of the test set alignments in Table 2.

For each test sets of each language pair, we released the abstracts in the source language and kept the ones in the target language for the both the automatic and manual evaluations, the so-called "reference translations". For instance, for the test set for de/en, we released the abstracts in German to the participants during the evaluation period and kept the ones in English for the evaluation.

2.2 Terminology

For the terminology dataset, a total of 6624 terms in English were manually translated to Spanish by domain experts. The terms were extracted from the scientific literature using the DNorm (Leaman et al., 2013) Named Entity Recognition and medical glossaries.

3 Baselines

Baseline 1: Marian NMT
This represents a low-experience, minimal effort submission based on current methods. We develop "baseline1" using the tutorial written for the MT Marathon 2018 Labs[9] and the Marian NMT framework (Junczys-Dowmunt et al., 2018).

As training data we used the UFAL medical corpus (UFA), and as validation data we used Khresmoi (Dušek et al., 2017). The Khresmoi data did not overlap with the UFAL corpus, despite being mentioned as one of the sources. The UFAL corpus was filtered to remove lower quality data. Specifically, we removed the "Subtitles" subset, as it is of lower quality than the rest, less medically oriented (if at all), and contains dialogue rather than narrative. Two of the targeted languages, Portuguese and Chinese, are not present in UFAL. For Portuguese we therefore trained our model on the Scielo corpus (Neves et al., 2016) and tested on the Brazilian thesis corpus (Soares et al., 2018b). For Chinese we used the United Nations Parallel Corpus (Ziemski et al., 2016).

The data was preprocessed using standard tools from the Moses toolkit (Koehn et al., 2007): tokenisation, cleaning of training data and truecasing. Subword segmentation (Sennrich et al., 2015) was then trained jointly over both source and target languages and applied using FastBPE.[10] The number of merge operations for BPE was set to 85000.

The models trained were shallow RNN encoder-decoders.[11] They were trained on a GTX 1080 Ti with 8 GPUs. Validation using cross-entropy and BLEU was performed every 10,000 updates, and models were trained until there was no improvement on either metric for 5 consecutive updates. Training of a single model took approximately 2 days.

Discussion. Compared to the traditional domain of news translation, biomedical MT poses additional challenges; biomedical texts contain a large amount of specialised, in-domain vocabulary, and in-domain training data is less readily available.

Baselines 2 and 3: OpenNMT
We also provide two additional baselines trained using OpenNMT-py (Klein et al., 2017)[12], one with a small network size, and a second one with a higher number of hidden units. The data used

[9] https://marian-nmt.github.io/examples/mtm2018-labs

[10] https://github.com/glample/fastBPE

[11] 1 encoder layer, 1 decoder layer, both with with GRU cells, embedding dimension of 512, hidden state of dimension 1024, using layer normalization, implemented using Marian NMT and trained using the Adam optimizer.

[12] http://opennmt.net/OpenNMT-py/

Language	OK	Source>Target	Target>Source	Overlap	No Align.	Total
de/en	808 (67.8%)	69 (5.8%)	126 (10.6%)	42 (3.5%)	147 (12.3%)	1192
es/en	825 (78.6%)	33 (3.1%)	67 (6.4%)	28 (2.7%)	96 (9.1%)	1049
fr/en	857 (82.6%)	21 (2.0%)	64 (6.2%)	9 (0.9%)	87 (8.4%)	1038
pt/en	833 (78.9%)	31 (2.9%)	77 (7.3%)	7 (0.7%)	107 (10.1%)	1055
zh/en	469 (84.4%)	53 (9.5%)	12 (2.2%)	5 (0.9%)	17 (3.1%)	556

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 of sentences corresponds to the 100 documents that constitute the two test sets (one for each language direction).

for training was the Medline abstracts corpora. We trained these two baselines using the following parameters:

- 2-layer LSTM for both the encoder and decoder (300 and 500 hidden units)
- Vocabulary size: 32,000
- Training steps: 100,000
- Batch size: 64
- Optimization: SGD
- Dropout: 0.3
- Embedding size: 500

The models were trained on a PC with Intel Xeon E-2124 processor and NVIDIA GeForce GTX 1060 GPU and are available for download.[13]

4 Teams and Systems

This year, the task attracted the participation of 11 teams from six countries (China, Germany, Japan, Pakistan, Spain and United Kingdom) from two continents. As opposed to previous years, no team from the Americas participated in the task. We list the teams and their affiliation (where available) in Table 3. We received a total of 59 run submissions as presented in Table 4.

System descriptions were solicited by email from the participating teams in the form of a system paper and a summary paragraph. Below we provide a short description of the systems for which a corresponding paper is available or for which we received a description from the participants. Two teams ('peace' and 'Radiant') did not provide system descriptions.

Table 5 provides an overview of the methods, implementations and training corpora used by the participants. While two teams used the statistical machine translation toolkit Moses (MT-UOC-UPF and UHH-DS), the most popular translation method relied on neural networks and the transformer architecture.

ARC (Wei et al., 2019). The ARC team's systems were based on the Transformer-big architecture (Vaswani et al., 2017). The relied on both general (news translation task, OPUS, UM, Wikipedia) and in-domain (EMEA, UFAL, Medline) corpora. For en/zh, they also used in-house training data. In order to improve the overall training data quality, they filtered noisy and misaligned data, and to improve vocabulary coverage they trained their subword segmentation model on the BERT multilingual vocabulary. They experimented with over 20 different models with various combinations of training data and settings and chose the best ones when submitting their runs.

BCS (Soares and Krallinger, 2019). The team's systems were also based on the Transformer-big architecture, which were trained using the OpenNMT-py toolkit. They relied on resources from both the general domain (books corpus), as well as from the biomedical domain, such as parallel terminologies from UMLS and various corpora (Scielo, UFAL medical, EMEA, theses and dissertations abstracts, and the Virtual Health Library).

KU. The KU team's systems were based on the Transformer-big architecture, trained using the Tensor2Tensor toolkit (Vaswani et al., 2018). Training data was carefully cleaned to remove encoding errors, bad translations, etc. They did not perform standard ensemble translation, but obtained a small BLEU improvement by taking a "majority vote" on the final translations for different checkpoints.

MT-UOC-UPF. The MT-UOC-UPF team's systems were deep RNN-based encoder-decoder models with attention, trained using Marian (and with layer normalisation, tied embeddings and

[13] 10.6084/m9.figshare.8094119

Team ID	Institution
ARC	Huawei Technologies (China),
BSC	Barcelona Supercomputing Center (Spain)
KU	Kyoto University (Japan)
MT-UOC-UPF	Universitat Oberta de Catalunya (Spain)
NRPU	Fatima Jinnah Women University (Pakistan), Manchester Metropolitan University (UK)
OOM	Beijing Atman Technology Co. Ltd. (China)
peace	(unknown)
Radiant	Harbin Institute of Technology (China)
Talp_upc	Universitat Politècnica de Catalunya (Spain)
UCAM	University of Cambridge (UK)
UHH-DS	University of Hamburg (Germany)

Table 3: List of the participating teams.

Teams	de/en	en/de	en/es	en/fr	en/pt	en/zh	es/en	fr/en	pt/en	zh/en	Total
ARC	M3	M3		M3		M3		M3		M3	18
BSC			M1		M1		M1		M1		4
KU				M1						M1	2
MT-UOC-UPF			M1T1				M1				3
NRPU				M1				M1			2
OOM						M2				M2	4
peace						M1				M1	2
Radiant						M3					3
Talpc_upc			M3				M3				6
UCAM	M3	M3	M3				M3				12
UHH-DS							M3				3
Total	6	6	9	5	1	9	11	4	1	7	59

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

Teams	MT method	Package, library or system	Training corpus
ARC	NMT	Transformer-big architecture	**general:** news translation task, OPUS, UM, Wikipedia; **in-domain:** EMEA, MEDLINE, UFAL
BSC	NMT	Transformer-big, OpenNMT-py	**general:** books corpus; **in-domain:** EMEA, Scielo, UFAL, UMLS, theses and dissertations abstracts, and the Virtual Health Library
KU	NMT	Transformer-big architecture, Tensor2Tensor toolkit	**in-domain**
MT-UOC-UPF	SMT, NMT	Moses, RNN-based Marian NMT	**in-domain**
NRPU	NMT	OpenNMT-py, transfer learning	**general:** News-Commentary; **in-domain:** EMEA, MEDLINE, Scielo, UFAL
OOM	NMT	Transformer architecture	**general** and **in-domain:** MedRA
peace	NA	NA	NA
Radiant	NA	NA	NA
Talpc_upc	NMT	Transformer architecture, BabelNet dictionary	**in-domain:** MEDLINE
UCAM	NMT	Transformer-big architecture, Tensor2Tensor toolkit, transfer learning	**general:** news translation task; **in-domain:** MEDLINE, Scielo, UFAL
UHH-DS	SMT	Moses	**in-domain:** biomedical task 2018 corpus

Table 5: Overview of the methods implemented by each team. We report the general translation method, specific package, library or implementation used and training corpus used. The letters "NA" are used when this information was not available at the time of writing.

residual connectors). The systems were trained with several medical corpora and glossaries. For the terminology translation task, they trained a Moses system using the same corpus as for the Marian NMT system. The translation table was queried for the English terms and when they were not found, they were translated using the NMT system if all subwords in the term were known and with the SMT Moses system if not.

NRPU (Noor-e-Hira et al., 2019). The NRPU team applied transfer learning and selective data training to build NMT systems. The goal of their approach is to mine biomedical data from general domain corpus and show its efficacy for the biomedical domain. The books corpus was used as the main out-of-domain corpus. News-Commentary (NC) (Tiedemann, 2012) was used as general domain corpus to perform information retrieval for selective data selection. The data selection procedure was performed as reported in Abdul-Rauf et al. (2016). In-domain MEDLINE titles were used as queries to retrieve biomedical related sentences from the general domain NC corpus. They had a total of 627,576 queries for data selection. Top n ($1<n<10$) relevant sentences were ranked against each query. The data selection process was done on both French and English.

OOM. Their system was based on the Transformer architecture trained on various parallel and monolingual corpora from in-domain and out-of-domain corpora. In the fine-tuning phase, the models were first tuned with the in-domain data and then fine-tuned with a curriculum learning mechanism for several rounds. Several model instances were ensembled to generate the translation candidates followed by a re-ranking model to select the best one. In addition to the standard sentences used in the training, terminological resources such as MedDRA were used as a constraint in the decoding phase to keep translation accuracy and consistency of key words.

Talp_upc (Pio Carrino et al., 2019). The Talp_upc team's submission was based on a Transformer and on the BabelNet multilingual semantic dictionary (Navigli and Ponzetto, 2012). From the Medline training data, they extracted a list of biomedical terms. They proposed *bpe-terms segmentation*, which consists of segmenting sentences as a mixture of subwords and term tokens in order to take into account domain-specific terms. They experimented with three systems: (i) terminology-aware segmentation (run2 for es/en and run2 for en/es), (ii) terminology-aware segmentation with a word-level domain feature (run3 for es/en and run1 for en/es), and (iii) terminology-aware segmentation, shared source and target vocabularies and shared encoder-decoder embedding weights (run1 for es/en and run3 for en/es).

UCAM (Saunders et al., 2019). The UCAM team relied on transfer learning and used the Tensor2Tensor implementation of the Transformer model. For each language pair, they used the same training data in both directions. Regarding training data, for en/de and de/en, they reused general domain models trained on the WMT19 news data and biomedical data (UFAl and Medline). For es/en and en/es, they trained on Scielo, UFAL, and Medline. Their three runs use the following: (i) the best single system trained on biomedical data, (ii) a uniform ensemble of models on two en/de and three es/en domains, and (iii) an ensemble with Bayesian Interpolation.

UHH-DS. The team submitted three runs for the Spanish-English language pair. Their SMT systems were developed using the Moses toolkit (Koehn et al., 2007) and trained on the same data as their submission from last year. Data selection was used to sub-sample two general domain corpora using a ratio of 50% sentences. Detailed descriptions of the methods are presented in (Duma and Menzel, 2016a) (run 1), (Duma and Menzel, 2016b) (run2) and (Duma and Menzel) (run 3). The first two methods rely on Paragraph Vector (Le and Mikolov, 2014) for sentence representation and scoring formulas based on the cosine similarity, and the third method focuses on the relative differences between term frequencies. All methods are unsupervised and produce fast results.

5 Automatic Evaluation

For each language pair, we compared the submitted translations to the reference translations. BLEU scores were calculated using the MULTI-EVAL tool and tokenization as provided in Moses. For Chinese, character-level tokenization was used via a minor modification to the tool. Although an ideal tokenization would take into account that Chinese words consist of a varying number of

characters, achieving such an ideal tokenization requires a sophisticated dictionary (Chang et al., 2008) – including biomedical terms – and is beyond the scope of this shared task. Further, using character-level tokenization for BLEU purposes is in accordance with current practice (Wang et al., 2018; Xu and Carpuat, 2018).

Table 6 shows BLEU scores for all language pairs when considering all sentences in our test sets. Table 7 only considers the sentences that have been manually classified as being correctly aligned (cf. Section 2). As expected, certain results improve considerably (by more than 10 BLEU points) when only considering the sentences that are correctly aligned.

Most teams outperformed the three baselines, except the NRPU team's submissions for en/fr and fr/en. Baseline1, trained using Marian NMT, obtained results not far behind the best performing team, while the two other baselines were not very competitive. We rank the various runs according to the results that they obtained followed by a short discussion of the results with regard to the methods that they used.

- de/en: baseline2,3 < baseline1 < UCAM, ARC
- en/de: baseline2,3 < baseline1 < UCAM, ARC
- es/en: baseline2,3 < baseline1 < UHH-DS < MT-UOC-UPF < BSC, Talp_upc runs2,3 < Talp_upc run1 < UCAM
- en/es: baseline2,3 < baseline1 < MT-UOC-UPF < Talp_upc, BSC < UCAM
- en/fr: baseline2,3 < NRPU < baseline1, KU < ARC runs2,3 < ARC run1
- fr/en: baseline2,3 < NRPU < baseline1 < ARC
- pt/en: baseline2,3 < baseline1 < BSC
- en/pt: baseline2,3 < baseline1 < BSC
- zh/en: baseline1 < peace < KU < ARC < OOM
- en/zh: Radiant < peace < ARC < OOM

de/en. All submitted runs from both ARC and UCAM teams outperformed our three baselines. The runs from ARC were slightly superior to those from UCAM. Both teams used Transformer models but the ARC also used BERT multilingual embeddings. We observed no significant difference between the submissions from team ARC but runs based on the ensemble of models from team UCAM (i.e. runs 2 and 3) obtained a higher score than their single best systems.

en/de. Results were similar to those for en/de: the runs from team ARC outperformed the runs from team UCAM. Similarly, we observed no difference between the runs from team ARC and slightly higher scores for the runs based on ensemble systems from team UCAM.

es/en. All submitted runs outperformed our baselines. The best performing systems from the Talp_upc, UCAM, and BSC teams were Transformer models, the one based on Marian NMT from the MT-UOC-UPF team, an finally the SMT Moses systems from UHH-DS. We did not observe significant differences between the various runs from single teams, except for run1 from Talp_upc team (terminology-aware segmentation, shared source and target vocabularies and shared encoder-decoder embedding weights), which outperformed their other two runs.

en/es. All submitted runs outperformed our baselines. As opposed to results for en/es, the Transformer system from the UCAM team slightly outperformed the one developed by the Talp_upc team, which obtained a similar performance to the OpenNMT system from the BSC team.

fr/en. Baselines 2 and 3 were outperformed by all submitted runs, whereas baseline 1, which is trained using Marian, was only outperformed by team ARC, whose system uses the Transformer model. We observed no significant difference between the three runs from the ARC team.

en/fr. Similar to fr/en, baselines 2 and 3 were outperformed by all submitted runs, while baseline 1 was similar to the run from the KU team, which uses the Transformer model. All runs from the ARC team outperformed our baseline 1. Run1 from the ARC performed significantly better than the other two runs, although details about the difference between the runs do not seem to be available.

pt/en. The run from the BSC team based on OpenNMT performed slightly better than baseline 1. However, their performance was far superior to baselines 2 and 3, which were also trained using OpenNMT but only trained on the Medline training data.

en/pt. Results for en/pt from the BSC were almost 10 points higher than the ones for pt/en. The run from the BSC team based on OpenNMT outperfomed with some difference the baseline based on Marian NMT, maybe because of the many resources that the team trained its system on. Further, they were much superior to the baselines 2 and 3 also based on OpenNMT but only trained on the Medline training data.

zh/en. All submitted runs outperformed the only baseline that we prepared. The three best-performing teams's submissions were Transformer models. The system developed by the OOM team slightly outperformed ARC's submission. Little difference in the results for the runs for the two teams was observed. A significant difference, however, was observed between results from the ARC and OOM teams and the Transformer system of the KU team.

en/zh. The Transformer-based system from team OOM significantly outperformed the transformer systems of team ARC. The latter had a similar performance to the runs for the other two teams (Radiant and peace) for which we do not know the details.

Table 8 presents the results of the automatic evaluation of the terminology test set. The evaluation considered the accuracy of translation (on lower-cased terms), rather than BLEU. The choice of accuracy was due to the fact that the terms are usually very short and having at least one different word from the reference translation can lead to a complete different meaning.

6 Manual Evaluation

For the Medline test sets, we performed manual evaluation of the primary runs, as identified by the participants, for all teams and language pairs. We carried out pairwise comparisons of translations taken either from a sample of the translations from the selected primary runs or the reference translations. Specifically, sets of translation pairs, consisting of either two automatic translations for a given sentence (derived from submitted results), or one automatic translation and the reference translation for a sentence, were prepared for evaluation. Table 9 presents the primary runs that we considered from each team. We performed a total of 62 validations of pairwise datasets.

We relied on human validators who were native speakers of the target languages and who were either members of the participating teams or colleagues from the research community. We also preferred to use validators who were familiar enough with the source language so that the original text could be consulted in case of questions about the translations, and for most language pairs this was the case.

We carried out the so-called 3-way ranking task in our installation of the Appraise tool (Federmann, 2010).[14]. For each pairwise dataset, we checked a total of 100 randomly-chosen sentence pairs. The validation consisted of reading the two translation sentences (A and B) and choosing one of the options listed below:

- A<B: the quality of translation B is higher than translation A;
- A=B: both translations have similar quality;
- A>B: the quality of translation A was higher than translation B;
- Flag error: the translations do not seem to come from the same source sentence, probably due to errors in the corpus alignment.

Table 10 summarizes the manual evaluation for the Medline test sets. We did not perform manual evaluation for any of our baselines. We ranked the runs and reference translations among themselves based on the number of times that one validation was carried out by the evaluators. When the superiority of a team (or reference translation) over another team was not very clear, we decided to put both of them together in a block without the "lower than" sign (<) between them. However, in these situations, the items are listed in ascending order of possible superiority in relation to the others. The various runs were ranked as listed below:

- de/en: reference, ARC, UCAM
- en/de: UCAM < ARC < reference

[14] https://github.com/cfedermann/Appraise

Teams and Runs	de/en	en/de	es/en	en/es	en/fr	fr/en	pt/en	en/pt	zh/en	en/zh
ARC-run1	0.2871	0.2789	-	-	0.3995	0.3551	-	-	0.3007	0.3547
ARC-run2	0.2879	0.2786	-	-	0.3667	0.3551	-	-	0.3005	0.3547
ARC-run3	0.2882*	0.2785*	-	-	0.3619*	0.3556*	-	-	0.3005*	0.3547*
BSC-run1	-	-	0.3769*	0.4421*	-	-	0.3990*	0.4811*	-	-
KU-run1	-	-	-	-	0.3114*	-	-	-	0.2489*	-
MT-UOC-UPF-run1	-	-	0.3659*	0.3974*	-	-	-	-	-	-
NRPU-run1	-	-	-	-	0.1587*	0.1972*	-	-	-	-
OOM-run1	-	-	-	-	-	-	-	-	0.3413	0.4234
OOM-run2	-	-	-	-	-	-	-	-	0.3413*	0.4234*
peace-run1	-	-	-	-	-	-	-	-	0.2266*	0.3379*
Radiant-run1	-	-	-	-	-	-	-	-	-	0.3266
Radiant-run2	-	-	-	-	-	-	-	-	-	0.3265
Radiant-run3	-	-	-	-	-	-	-	-	-	0.3294*
Talp_upc-run1	-	-	0.3941	0.4301*	-	-	-	-	-	-
Talp_upc-run2	-	-	0.3792*	0.4340	-	-	-	-	-	-
Talp_upc-run3	-	-	0.3721	0.4392	-	-	-	-	-	-
UCAM-run1	0.2741	0.2651	0.4241	0.4492	-	-	-	-	-	-
UCAM-run2	0.2863	0.2716	0.4303	0.4539	-	-	-	-	-	-
UCAM-run3	0.2850*	0.2641*	0.4290*	0.4558*	-	-	-	-	-	-
UHH-DS-run1	-	-	0.3561*	-	-	-	-	-	-	-
UHH-DS-run2	-	-	0.3585	-	-	-	-	-	-	-
UHH-DS-run3	-	-	0.3586	-	-	-	-	-	-	-
baseline1	0.2473	0.2202	0.3306	0.3722	0.3056	0.2927	0.3812	0.4115	0.1519	-
baseline2	0.0954	0.0347	0.2373	0.0614	0.0211	0.1406	0.2280	0.2264	-	-
baseline3	0.0962	0.0367	0.2373	0.0614	0.0221	0.1572	0.2394	0.2328	-	-

Table 6: BLEU scores when considering all sentences in the test sets. Runs are presented in alphabetical order of the team's name, while the baseline results are shown at the bottom of the table. * indicates the primary run, as indicated by the participants, in the case of multiple runs.

Teams and Runs	de/en	en/de	es/en	en/es	en/fr	fr/en	pt/en	en/pt	zh/en	en/zh
ARC-run1	0.3866	0.3539	-	-	0.4241	0.3818	-	-	0.3215	0.3709
ARC-run2	0.3880	0.3528	-	-	0.3889	0.3818	-	-	0.3216	0.3709
ARC-run3	0.3884*	0.3526*	-	-	0.3829*	0.3824*	-	-	0.3216*	0.3709*
BSC-run1	-	-	0.4356*	0.4701*	-	-	0.4617*	0.4951*	-	-
KU-run1	-	-	-	-	0.3292*	-	-	-	0.2716*	-
MT-UOC-UPF-run1	-	-	0.4159*	0.4219*	-	-	-	-	-	-
NRPU-run1	-	-	-	-	0.1745*	0.2105*	-	-	-	-
OOM-run1	-	-	-	-	-	-	-	-	0.3561	0.4392
OOM-run2	-	-	-	-	-	-	-	-	0.3561*	0.4392*
peace-run1	-	-	-	-	-	-	-	-	0.2518*	0.3508*
Radiant-run1	-	-	-	-	-	-	-	-	-	0.3405
Radiant-run2	-	-	-	-	-	-	-	-	-	0.3416
Radiant-run3	-	-	-	-	-	-	-	-	-	0.3424*
Talp_upc-run1	-	-	0.4509	0.4568*	-	-	-	-	-	-
Talp_upc-run2	-	-	0.4355*	0.4609	-	-	-	-	-	-
Talp_upc-run3	-	-	0.4270	0.4683	-	-	-	-	-	-
UCAM-run1	0.3669	0.3328	0.4770	0.4834	-	-	-	-	-	-
UCAM-run2	0.3807	0.3469	0.4833	0.4891	-	-	-	-	-	-
UCAM-run3	0.3799*	0.3379*	0.4811*	0.4896*	-	-	-	-	-	-
UHH-DS-run1	-	-	0.3969*	-	-	-	-	-	-	-
UHH-DS-run2	-	-	0.3999	-	-	-	-	-	-	-
UHH-DS-run3	-	-	0.3997	-	-	-	-	-	-	-
baseline1	0.3277	0.2806	0.3765	0.4037	0.3253	0.2989	0.4298	0.4275	0.1667	-
baseline2	0.1250	0.0410	0.2724	0.0633	0.0228	0.1553	0.2666	0.2284	-	-
baseline3	0.1287	0.0436	0.2724	0.0633	0.0236	0.1730	0.2727	0.2345	-	-

Table 7: BLEU scores when considering only the correctly aligned sentences in the test sets. Runs are presented in alphabetical order of the team's name, while the baseline results are shown at the bottom of the table. * indicates the primary run, as indicated by the participants, in the case of multiple runs.

Teams	Runs	en/es
MT-UOC-UPF	1	47.55

Table 8: Accuracy results for the terminology test set.

- en/es: reference, MT-UOC-UPF < BSC, Talp_upc, UCAM
- en/fr: NRPU < KU < ARC, reference
- en/pt: reference, BSC
- en/zh: no possible ranking
- es/en: UHH-DS < MT-UOC-UPF < BSC, UCAM < reference, Talp_upc
- fr/en: NRPU < reference < ARC
- pt/en: BSC, reference
- zh/en: KU < ARC, peace < reference, OOM

The ranks for the manual validation were usually consistent with the ones that we obtained for the automatic validation. We discuss differences that we found in the discussion of the results for each language pair below.

de/en. The reference translations and the runs from teams ARC and UCAM were of similar quality and we did not observe huge differences between them. For this reason, we have grouped them into a single block, ordering them according to increasing performance. The UCAM team's submission was seen to be marginally better than the reference translations (33 vs. 23). We did not observe any differences in the respective order of teams compared to that of the automatic evaluation.

en/de. The reference translation was clearly superior to the runs from the ARC and UCAM teams (41 vs. 19, and 44 vs. 16, respectively). The translations from the ARC submission were more frequently judged better than the ones from the UCAM team (37 vs. 16). While we found no significant difference in the BLEU scores for teams ARC and UCAM, the manual evaluation showed that translations from team ARC were of superior quality to those of team UCAM.

en/es. The runs from the MT-UOC-UPF and BSC teams were judged as of similar quality to the reference translations, while the ones from Talp_upc and UCAM were deemed superior to the reference translations. The manual validation did not indicate much difference between runs from teams BSC, Talp_upc and UCAM. The ranking of the teams did not change significantly between that of the automatic evaluation.

en/fr. The reference translations were clearly superior to the runs from the KU and the NRPU teams, however, they were found only marginally superior to the ARC run. We therefore decided to put the ARC runs and reference translations in a single block. As for the comparison of the ARC runs to the KU and NRPU runs, superiority of ARC was higher when compared to the NRPU team (82 vs. 2) than for team KU (42 vs. 21). Indeed, the translations from the KU team were validated as far superior (73 vs. 9) to team NRPU. We did not observe any differences in the ranking of teams with respect to the automatic evaluation.

en/zh. We could not rank the runs from the various teams because of inconsistencies when comparing results from the various pairwise validations. For instance, the translations from the OOM team were judged better than the the reference translations, and the latter better that the ones from the ARC team. However, the translation from the ARC team were considered better than the ones from the OOM team. We also found differences in the rankings found in the automatic validation. For instance, the team that obtained the lowest BLEU scores (peace), had their translation judged to be as good as the ones from the Radiant and OOM teams, two of the teams that obtained high BLEU scores.

en/pt. The translations from the BSC team were validated as slightly superior (29 vs. 25) to the reference translations. We therefore grouped both of them in a single block.

es/en. The reference translations were judged as of similar quality to the ones from the Talp_upc teams, followed by the translations from the BSC and UCAM teams. The only difference to the ranking from the automatic evaluation was that the runs from the Talp_upc were considered better than those from the UCAM team while the latter obtained a higher BLEU score.

Teams	de/en	en/de	en/es	en/fr	en/pt	en/zh	es/en	fr/en	pt/en	zh/en	Total
ARC	run3	run3		run3		run3		run3		run3	6
BSC			run1		run1		run1		run1		4
KU				run1						run1	2
MT-UOC-UPF			run1				run1				2
NRPU				run1				run1			2
OOM						run2				run2	2
peace						run1				run1	2
Radiant						run3					1
Talpc_upc			run1				run2				2
UCAM	run3	run3	run3				run3				4
UHH-DS							run1				1
Total	2	2	4	3	1	4	5	2	1	4	28
Pairwise	3	3	10	6	1	10	15	3	1	10	62

Table 9: Overview of the primary runs that were considered for manual validation. The last columns shows the number of runs that we validated for each team. The last rows in the tables show the total number of runs and of pairwise combinations between runs and the reference translations.

fr/en. The reference translations were consistently validated as superior to the one from team NRPU's submissions, whereas the ones from team ARC were judged to be better than the reference translations.

pt/en. The reference translations were validated as slightly superior (29 vs. 24) to the ones from team BSC. Therefore, we grouped both of them in a single block.

zh/en. Only the translation from the OOM team, the runs that obtained the highest BLEU scores, were judged as of similar quality to the reference translations. The only main difference compared to the ranking from the automatic translation was with regard to team peace's submission, which obtained the lowest BLEU score, but for which the translations were ranked higher than the ones from the KU team and of similar quality to the ARC team according to the manual evaluation.

7 Discussion

In this section we discuss important topics related to the shared task, such as a short analysis of best performing methods, lack of sufficient resources for some language pairs and the quality of the test sets and the submitted translations.

7.1 Analysis of results and methods

Across all language pairs, the best performing runs were those based on the Transformer architecture trained on as much data as possible from the general and biomedical domain (cf. the submissions by the ARC, Talp_upc, and UCAM teams). Ensembled runs tended to perform well and generally outperformed using the single best system (cf. OOM, Talp_upc, and UCAM).

Differences in the amount of training data available across languages appeared to have a direct impact on translation quality. The Scielo and Medline corpora are larger for es/en and en/es than for the other languages, which is reflected in the BLEU scores. For example, results for team UCAM were more than 10 points higher for es/en and en/es than for de/en and en/de, results which were mirrored for baseline 1.

Regarding zh/en and en/zh for which we do not yet provide any training data, results were inferior to the best-performing language pairs (es/en and en/es), but still surprisingly good. However the best-performing teams trained on additional in-house data (cf. ARC's submission), which was not available to the community.

We compared results for this year's shared task in comparison to the previous year's (Neves et al., 2018). The addition of the Medline training data this year resulted in an improvement for en/de (from 24.30 to almost 28.00), but not for de/en. Similarly, we observed no real improvement for es/en and en/es, the highest BLEU scores for both remained in the range of 43-45 points. However, a considerably improvement occurred for en/fr, whose scores increased from almost 25 to almost 40 points, and for fr/en from almost 27 to around 35 points. Finally, the scores for en/pt showed an improvement from 43 to 49 points, while the scores remained constant for pt/en on 46 points.

In the shared task that we organized last year (Neves et al., 2018), for the first time certain runs outperformed the reference translations in the

Languages	Runs (A vs. B)	Total	A>B	A=B	A<B
de/en	reference vs. ARC	94	31	30	33
	reference vs. UCAM	93	23	37	33
	ARC vs. UCAM	100	20	60	20
en/de	reference vs. ARC	92	41	32	19
	reference vs. UCAM	92	44	32	16
	ARC vs. UCAM	100	37	47	16
en/es	reference vs. BSC	100	10	78	12
	reference vs. MT-UOC-UPF	100	25	49	26
	reference vs. Talp_upc	100	7	74	19
	reference vs. UCAM	100	18	62	28
	BSC vs. MT-UOC-UPF	100	26	59	15
	BSC vs. Talp_upc	100	9	80	11
	BSC vs. UCAM	100	9	86	5
	MT-UOC-UPF vs. Talp_upc	98	6	77	15
	MT-UOC-UPF vs. UCAM	100	6	75	19
	Talp_upc vs. UCAM	100	11	82	7
en/fr	reference vs. ARC	98	36	34	28
	reference vs. KU	98	61	21	16
	reference vs. NRPU	99	79	18	2
	ARC vs. KU	100	42	37	21
	ARC vs. NRPU	100	86	12	2
	KU vs. NRPU	99	73	17	9
en/zh	reference vs. ARC	95	55	12	28
	reference vs. OOM	100	28	28	44
	reference vs. peace	93	50	18	25
	reference vs. Radiant	99	24	14	61
	ARC vs. OOM	96	52	7	37
	ARC vs. peace	96	52	7	37
	ARC vs. Radiant	93	45	6	42
	OOM vs. peace	100	33	38	29
	OOM vs. Radiant	100	68	16	16
	peace vs. Radiant	98	43	7	48
en/pt	reference vs. BSC	99	25	45	29
es/en	reference vs. BSC	98	40	30	28
	reference vs. MT-UOC-UPF	90	36	36	10
	reference vs. Talp_upc	95	27	42	26
	reference vs. UCAM	99	30	45	24
	reference vs. UHH-DS	96	55	33	8
	BSC vs. MT-UOC-UPF	97	32	39	26
	BSC vs. Talp_upc	100	19	43	38
	BSC vs. UCAM	99	29	48	22
	BSC vs. UHH-DS	100	55	29	16
	MT-UOC-UPF vs. Talp_upc	95	15	46	34
	MT-UOC-UPF vs. UCAM	100	24	35	41
	MT-UOC-UPF vs. UHH-DS	100	51	36	13
	Talp_upc vs. UCAM	100	33	42	25
	Talp_upc vs. UHH-DS	100	55	35	10
	UCAM vs. UHH-DS	98	54	34	10
fr/en	reference vs. ARC	96	23	32	41
	reference vs. NRPU	95	72	20	3
	ARC vs. NRPU	99	80	19	0
pt/en	reference vs. BSC	96	29	43	24
zh/en	reference vs. ARC	100	47	29	24
	reference vs. KU	100	36	37	27
	reference vs. OOM	100	12	43	12
	reference vs. peace	100	33	32	25
	ARC vs. KU	100	36	44	20
	ARC vs. OOM	100	13	41	46
	ARC vs. peace	100	31	38	31
	KU vs. OOM	100	9	40	51
	KU vs. peace	100	25	42	33
	OOM vs. peace	100	49	45	6

Table 10: Results for the manual validation for the Medline test sets. Values are absolute numbers (not percentages). They might not sum up to 100 due to the skipped sentences.

Pair	=	>
de/en	ARC, UCAM	-
en/de	-	-
en/es	MT-UOC-UPF	BSC, Talp_upc, UCAM
en/fr	ARC	-
en/pt	BSC	-
en/zh	-	-
es/en	Talp_upc	-
fr/en	-	ARC
pt/en	BSC	-
zh/en	OOM	-

Table 11: List of teams with runs of a similar quality to the reference translations or that outperformed them.

manual validation (e.g. for en/es) or were of similar quality (e.g. de/en). This year there were more such cases (cf. Table 11), which confirms the improvements of the participating systems.

7.2 Quality of the test sets

To evaluate the quality of the MEDLINE test sets, we performed an evaluation of the sentence alignment using Appraise to classify sentence pairs between "OK", "Target > Source", "Source > Target" and "No Alignment". During this process, we also noted any observation on the quality of the reference translation. Of note for this dataset, the reference translation is produced by the original authors of the papers who are scientists with likely no training in translation and whose writing competence in the languages involved is unknown. We can make the hypothesis that the authors have acquired English as a second language while they have native or near-native competence in the non English language.

The quality of the alignment in the Medline test sets varied from as low as around 68% (for de/en and en/de) to as high as 84.4% (for zh/en and en/zh). Therefore, the rate of misaligned sentences did not vary much across the language pairs. Part of this problem was due to incorrectly considering the titles of the citations, when usually there is no translation for these available in Medline.

Some of the segments assessed as correctly aligned ("OK") sometimes exhibited sentence segmentation error that were similar in the two languages. For example, there were segments where pairs of sentences were aligned, instead of being split into two aligned segments.

Interestingly, except for zn/en and en/zh, we observed an average of twice more sentences classified as "Target > Source" than as "Source > Target". This might suggest that authors of the articles might have added more information in the English version of the article than in the version in the foreign language.

During our manual validation of the test sets (cf. Section 2), we identified the non-aligned sentences with the specific label 'No Alignment'. However, almost 1/3 of these not aligned sentences correspond to other issues: (a) misalignment between titles to nothing or something else; (b) misalignment of complete, different sentences (even though these were rather rare); and (c) misalignment of section+sentences wrongly aligned to only the section name in the other language. The latter was also sometimes classified as either "Target > Source" or "Source > Target". Regarding these two labels, i.e., "Target > Source" and "Source > Target", these were often utilized for the following situations: (a) section+sentence automatically aligned only to sentence (the opposite of the above); (b) reference to an entity (e.g. a disease), while referred only to the pronoun (e.g. it) in the other language; (c) mention of a particular information (e.g. a method or a time range) in one language, while not in the other language; and (d) the English version included notes in squared brackets which were not part of the foreign sentence.

We also identified problems in the reference translation when performing the manual validation. Some issues were related to the sentence splitting, for instance, p-values were often split, so that "(p=0.5)" would be split on the ".". In those cases, the preceding sentence ended in "...(p=0." and the next sentence started with "5) ...". Others were related to the content of the reference translations themselves, including non-literal translations that alter the meaning of the original sentence when out of context (Example 1), wrong translations (as in Example 2) and even poor formatting and punctuation.

(1) *Source:* Toutes **ces personnes**, et en particulier dans le monde du sport amateur...
Ref: **These athletes**, especially, the amateurs...
Correct: All of **these people**, especially in the amateur sports world...[15]

(2) *Source:* Les crises épileptiques sont imprévisibles et peuvent se produire

[15] Relevant parts of the translation are indicated in bold. The same holds for Example 2.

n'importe où.
Ref: Epileptic seizures occur with unpredictable frequency **in unexpected place**.
Correct: Epileptic seizures are unpredictable and can occur **anywhere**.

A further problem identified was the presence of very short sentences often formed of a single word (e.g. titles or listed items such as "Conclusions", "Objective", or "Clinical Case"), which are in general correctly translated. Including such items for evaluation could influence quality assessments, inflating the scores, since their translation is more similar to terminology translation rather than sentential translation.

7.3 Quality of the system translations

English (from Chinese). As the first year receiving submissions addressing the Chinese language, the overall quality of the translations was delightfully high. For an English sentence to offer the same level of fluency, the order of phrases is often different from those in the source Chinese sentence. Many of the submitted translations successfully captured this behavior, as in the example below.

在健康风险和生理及心理自我调节能力评估讨论的背景下解读 HRV 节律。
(Order of terms: health risk, physiological and psychological self-regulation, interpretation, HRV rhythms.)
– Source

Interpretation of heart rate variability rhythms in the context of health risk and physiological and psychological self-regulatory capacity assessment is discussed.
– Reference translation

HRV rhythms are interpreted in the context of health risks and assessment of physiological and psychological self-regulation.
– Translation

Errors that disrupt the meaning of the translations most are incorrectly translated biomedical terms, presumably due to an inadequate Chinese biomedical dictionary. For instance, 人智医学 or *anthroposophic healthcare* was, based on the literal meanings of the individual Chinese characters making up the term, variably translated as *human intellectual healthcare*, *psychiatric care* and even *humane healthcare*. Other literal but incorrect translations include *horse's syndrome* for 马方综合征 (Mafran's syndrome) due to the 马 character (a horse), and *parasitic therapy* for 槲寄生疗法 (mistletoe therapy) due to 寄生 (parasitic). In some cases, such terms, which were presumably absent from the dictionary, were entirely omitted in the translations.

Improvements to the translations could be made in two areas. Firstly, singular and plural markings could be made consistent within one translated abstract. In Chinese, with very few exceptions, nouns are not inflected to indicate plurality. Hence where an earlier sentence in an abstract mentions, for instance, 两名患者 (*two patients*) and in a later sentence only 患者, a correct English translation should remain consistent with the plural *patients*, not the singular *patient*. Secondly, non-biomedical terms with connotations specific to scientific abstracts could be more precisely translated. For instance, beginning the final sentence in an abstract with 总之 would be better translated as *in conclusion* than *in general*.

English (from French). The overall translation quality was high for this language direction, and it was often difficult to distinguish between the MT output and the reference translation in terms of quality, in some cases indicative of the good quality of automatic translation, and in others of the presence of problems in the reference translations themselves.

An aspect that could have contributed to a translation being considered better or worse was the handling of complex noun phrases (e.g. *case monitoring* versus a prepositional phrase complement *monitoring of cases*). Whereas many prescriptivists would prescribe the noun compound variant, these were actually often perceived to be more natural and appropriate for academic or scientific writing.

Noun compound	**PP complement**
robust case monitoring	robust monitoring of cases
stool culture results	results of stool culture
treatment trajectories	trajectories of treatments

Table 12: Examples of equally grammatical noun compounds and prepositional phrase (PP) complements in the fr/en manually evaluated data.

English (from German). The quality of the translations from German to English was gen-

erally good. German sentences, which have a typically different structure and word order than English sentences, were usually re-arranged with conjunctions and subordinate clauses in proper written English. In a few cases, the greater context of the German corpus at hand appeared to influence the translation of the individual checked sentences, as additional information, which was not part of the original German sentence, was integrated into the English translation. For example:

> *Bei 3,6% war schon einmal eine psychosomatische Reha durchgeführt worden und dennoch vom Konsiliararzt eine Wiederholungsreha als sinnvoll erachtet.* ***Patienten, die bereits einmal in Reha waren sind kränker und haben mehr Fähigkeits- und Teilhabeprobleme.***
> *Von 35 Patienten, bei denen der Konsiliararzt die Neubeantragung einer psychosomatischen Rehabilitation empfahl, wurde bei 13 im Verlauf der folgenden 6 Monate ein Antrag gestellt.*
> – Source
>
> *Patients who had already been in inpatient rehabilitation in the past 5 years were more severely ill and had more severe participation problems.*
> – Translation

As the appraiser was blinded to the source of the translations, it was not possible to determine if such sentences were machine-translated or human-translated.

Pro-forms were also successfully used in the German to English translations, such as *sie* to *OCT*, referring to *optical coherence tomography*, and *In den aufgearbeiteten Fällen* to *In our cases*. These two examples make sense and appear to be correctly translated. However, other pro-forms were not, such as German *er* to English *he* instead of the gender-neutral pronoun *it*. German pronouns present a challenge for automated translation, as all nouns in the German language are assigned a gender, whereas in English, only persons are given gendered pronouns.

While most German words were correctly translated to their English equivalents, there were some interesting cases, ranging from completely off-topic to understandable yet odd equivalents. For example, the German word *Möpsen* (English: *pugs*) was incorrectly translated many times to *seagulls* or *cups*. *Konsiliararzt* (English: *consultant*) was translated to different terms but never correctly: *siliconist* or *silicone doctor*. Interestingly, the adverb *konsiliarärztlich* was correctly interpreted to describe a recommendation from a doctor in the English translation, but unfortunately this doctor was translated to be a *silicone doctor*:

> *Bei 64% der Patienten mit chronischen psychischen Erkrankungen war bislang keine psychosomatische Reha erfolgt und auch keine Indikation gegeben.* ***Bei 27% wurde bislang noch keine Rehamaßnahme durchgeführt, wurde jetzt aber konsiliarärztlich erstmals empfohlen.***
> *Bei 3,6% war schon einmal eine psychosomatische Reha durchgeführt worden und dennoch vom Konsiliararzt eine Wiederholungsreha als sinnvoll erachtet.*
> – Source
>
> *At 27%, no rehab has been performed yet, but has now been recommended for the first time by a silicone doctor.*
> – Translation

Improvements to automated translation could be made if translations of medical or technical words could be constrained to the context. When describing the torso of the human body, *Rumpf* was translated to the aviation term *fuselage* and *Säugezeit* was literally translated to *mammalian period* instead of *suckling period*. In peculiar yet comprehensible translations, the German *befragte Person* was translated to *repliers* instead of *respondents*. The English translation of *Lebensqualität* was mistaken as the phonetic *quality of live* instead of *quality of life*. On a positive note, the German false friend *evtl.* was indeed correctly translated to the English word *possible*. Some abbreviations were not even translated at all (*AÄ, OÄ, KP*), yet the procedure *Zementsakroplastie (ZSP)* was correctly constructed as *Cement Sacroplasty (CSP)* in English. *Vitien* (English: *cardiac defect*), which is actually Latin, was wrongly translated to *vials* or *vii*. Overall, German scientific and medical terms and abbrevi-

ations were occasionally difficult to translate correctly.

In a handful of examples, the English translations appeared to be too colloquial for a written scientific context. This includes phrases such as *so you always have to ask about it* and *but there are no studies on that* and using the verb *got* instead of *received*. From the appraiser's point of view, the origin of these phrases - automatic translation or manually curated gold standard - is not clear.

In a few cases, the English translations, despite being grammatically correct, altered the intended meaning of the original German sentence. The compound word *Teilhabebeeinträchtigungen* was wrongly translated to *partial impairment* instead of *participation impairment*. In another example, a long German sentence ending in *Antrag gestellt* was incorrectly interpreted to mean *received an application*. The same original text was further mistakenly interpreted in another translation to imply that the actual rehabilitation had been started, when in fact the German original indicated that only an application for rehabilitation had been initiated:

> *Patienten, die bereits einmal in Reha waren sind kränker und haben mehr Fähigkeits- und Teilhabeprobleme.* ***Von 35 Patienten, bei denen der Konsiliararzt die Neubeantragung einer psychosomatischen Rehabilitation empfahl, wurde bei 13 im Verlauf der folgenden 6 Monate ein Antrag gestellt.***
> *SCHLUSSFOLGERUNG*
> – Source

> *Of 35 patients in whom the consultant recommended the reapplication of psychosomatic rehabilitation, 13 received an application during the following 6 months.*
> – Translation A

> *In 13 out of 35 patients who got a recommendation for a new psychosomatic rehabilitation, this rehabilitation was initiated within 6 months after the consult.*
> – Translation B

> *Of the 35 patients in whom the silicone doctor recommended a new application for psychosomatic rehabilitation, 13 were applied for during the following 6 months.*
> – Translation C

In fact, Translation C was the most correct about the 13 patients, except the error that *Konsiliararzt* was translated as *silicone doctor*.

In a last example, the words *nachhaltigen Effekt* were translated to two different possibilities: *sustainable effect* (the fact that the effect is able to be sustained) and *sustained effect* (an effect that held continuously at a certain level). There is a subtle difference in meaning of these two English terms, whereas the German word (*nachhaltigen*) could used to describe both situations. This complicates a straight-forward translation because the correct interpretation heavily depends on the whole context of the matter:

> *Berufsgruppenbedingte Unterschiede im klinischen Alltag und individueller Karrierefortschritt üben einen Einfluss auf Art, Umsetzung und Wahrnehmung der Lehrtätigkeit aus.* ***Hinweise auf einen nachhaltigen Effekt ermutigen zur Fortsetzung und Weiterentwicklung des TTT-Konzepts.***
> *Er wurde in den letzten acht Jahren auf ähnliche Symptome untersucht.*
> – Source

> *Indications of a sustained effect encourage the continuation and further development of the TTT concept.*
> – Translation A

> *Indications of a sustainable effect encourage the continuation and further development of the TTT concept.*
> – Translation B

From the appraiser's point of view, it is not possible to ascertain the author's true meaning of *nachhaltigen* from these short excerpts.

English (from Spanish). The translations into English from Spanish were notably improved this year , and judgments were much more subtle in

many cases. There were still a few occurrences of untranslated words appearing in the translations, but far fewer than in previous years.

Lexical choice was often a differentiating factor between translations, e.g. *accomplish several goals* was preferred to *achieving various goals*.

Grammar differences were also visible, in particular for complex noun phrases, e.g. *creative alternatives management* vs. *creative management alternatives*.

Some differences in the translations hinged on treatment of acronyms; without further context (i.e., the expansion of the acronym) or specific domain knowledge it was sometimes difficult to decide which acronym should be preferred.

Reference translations were sometimes clearly identifiable due to including information from other parts of the text outside of the focus sentence, leaving out some details in the original, or completely rephrasing an idea; in general translations more faithful to the original sentence were preferred, as long as the translation was basically fine.

Sometimes neither translation being compared was ideal, and assessment came down to a judgment call. For instance, comparing the two translations A: *In the **double cerclage**, surgery time was shorter (average 38 minutes), and the range of motion showed improvement **since** the first month.* and B: *In the **cerclage double**, the time of surgery was shorter (average 38 minutes), and the range of motion demonstrated improvement **from** the first month.*, A has the more accurate grammar for *double cerclage*, but *from the first month* is more correctly expressed. In this case, B was picked because the error in the noun phrase is easier to compensate for.

Another such example was the translation of *Existen desigualdades de género en la provisión de cuidados informales a mayores dependientes en Gipuzkoa, mostrando las mujeres un mayor impacto en su salud y CVRS que los hombres.* as A: *There are gender inequalities in the provision of informal care to dependent older adults in Gipuzkoa, showing that women have a greater impact on their health and HRQOL than men.* and B: *Gender inequalities exist in the provision of informal care to elderly dependent in Gipuzkoa, showing women a greater impact on their health and HRQL than men.*. Both translations are imperfect, however A provides a better treatment of *mayores dependientes (the dependent elderly)* than B – although B is close, it requires a plural *dependents*. However, *showing that women* is not a natural way to express the relationship between the gender inequalities (*desigualdades de género*) in the first half of the sentence and the impact of women in the second half; a better translation would be *indicating that women* or *with women having*. On balance, though, translation A is overall more readable than B.

Some differences were only in relation to spacing, i.e. one translation included "patients,14%" while the other had "patients, 14%". This suggests the use of character-level modeling in the algorithms having occasional hiccups. One particularly problematic translation was *Univariate and **multivariateanalyses** were performed through a Multilayer **Perceptronnetwork** and a logistic regression model **EmpiricalBayesian** penalized type LASSO Elastic net*. On the flip side, these algorithms were sometimes able to correct spacing problems in the source text.

Chinese. The quality of translations from all four participating systems was very high, and the translations were generally fluent and accurate. When comparing the translations from the various systems, shorter sentences were typically highly similar, differing only in certain formulations. However, such differences could suffice to distinguish one translation as better than another, because a wording (e.g. 新努力 *new efforts*) more precisely captures the source (exactly *new efforts*) than alternative wordings (新进展 *new developments*). For longer sentences, more noticeable differences surfaced, particularly in different orderings of phrases. These orderings sometimes impacted the fluency of the translation, but in generally were merely different but valid arrangements of the same content.

In terms of serious errors, only in rare cases were phrases completely dropped in the translations. As for incorrect translation of biomedical terms, they occurred far less frequently in the en/zh direction than zh/en. One might hypothesize that the dictionary in the en/zh direction was more complete. However, the fact that translating into Chinese has the option of retaining the original term in English is also a contributing factor, which leads us to the next point.

Currently there is no consensus in how much of a technical term in English should be preserved

in the Chinese translation. Take *Functional electronic stimulation (FES)* in a source as an example. Valid translations in Chinese include having only the Chinese term (功能性电刺激); with the acronym (功能性电刺激 *(FES)*); as well as with full term plus acronym (功能性电刺激 *(Functional electronic stimulation, FES)*). Gene names, on the other hand, are uncontentiously retained in English (e.g. *AMP* and *CK2* α in source, reference, and submitted translations alike).

German. Compared to last year again in general translations were of very high quality. Only rarely we found untranslated bits from the source language, while automatic systems were mostly able to differentiate between sequences that should be translated or not (e.g. citations, links). The use of capitalization was correct in almost all cases. Therefore, the decision for a better translation was mostly based on the correct translation of technical terms, in general a more appropriate use of German words or word order.

Mostly usage of technical terms was decisive: grayscale ultrasound is *Schwarz-Weiß-Ultraschall* instead of *Graustufen-Ultraschall*, or similarly *mandibular advancement device* is a *Unterkieferprotrusionsschiene* instead of the rather word-by-word but wrong translation *mandibulären Fortschrittsgerät*. Other examples rather concern the appropiate use of German words. For instance, *disease attenuation* is rather a *Abschwächung* than a *Dämpfung* of a disease. It seems that automatic systems could not deal with more complex syntax such as coordination as in *tumor mass and symptom reduction*. Instead of *Tumormassenreduktion und Symptomlinderung*, the automatic translations did not identify the coordination structure and produces an incorrect (word-by-word) translation *Tumormasse und Symptomreduktion*.

Similar to last year, cases when automatic systems were judged better than the reference, the reference contained additional information or missed information while translation usually contained the complete content of the source sentence.

We were not able to define clear patterns for differences between the two automatic systems. However, ARC seems to be more capable of providing proper German syntax (e.g. *Steifigkeitsschwankungen* for *stiffness variation* or *Patienten mit Bauchspeicheldrüsenkrebs* for *pancreatic cancer patients* than UCAM. On the other hand, ARC seems to have difficulty identifying acronyms at the beginning of a sentence and did not keep them all capitalized. ARC even provided a false translation for *Sleep is ... unrefreshing* as *Schlaf ist ... erfrischend* instead of *nicht erfrischend*. UCAM did not show the last two issues.

French. Although the quality of the translations was generally uneven, some systems offered mostly fluent translations.

A number of errors were easily identified as untranslated segments, or repeated words. However, a category of serious errors occurred in otherwise fluent sentences where missense or erroneous information was introduced. This is the case for example when a significant piece of information is omitted in the translation: *We used inverse proportional weighting* translated by *Nous avons utilisé un facteur de pondération proportionnelle* (omission de *inverse*) or when numbers are substituted: *data from adolescents aged 15-18 years* translated by *données des adolescents âgés de 12 à 25 ans*. Arguably, in these cases, no translation would be preferable to a translation error that could easily go undetected.

One notable improvement over previous years was the processing of acronyms, which were often directly expanded or translated with suitable equivalents: for example, *long-lasting insecticidal nets (LLINs)* was translated by *moustiquaires imprégnées d'insecticide de longue durée (MILD)* or *moustiquaires imprégnées à longue durée d'action (MILDA)*. Further assessment should take context beyond a single sentence into account, so that the consistency of use of acronyms can be evaluated over a document. It can also be noted that in some cases, the context of a sentence is not enough to make an assessment. For example, the phrases *Elle survient le plus souvent...* ou *Il se développe le plus souvent...* could be acceptable translations for *It occurs most frequently...*, depending on the grammatical agreement between *Elle/Il* and the translation of the antecedent.

Portuguese. As shown in the results for manual validation (cf. table 6), the automatic translations for Portuguese were of very good quality and often with similar or higher quality as the reference translations. However, we still found some mistakes and issues. Similar to previous years, we still find some acronyms, words or phrases (e.g. Leo G. Reeder Award) that were not translated and remained in the English format. We also found

some small mistakes when referring to particular values or parameters from the study, usually between parenthesis. For instance, the passage "88% para T2-0,535 cm)" instead of the complete statement "88% para RM ponderada em T2 (viés = 0,52 cm2; p = 0,735)".

We identified few terms that were translated literally into Portuguese. For intance the term "scrub nurses" was translated into "enfermeiros esfregadores" instead of "enfermeiros/instrumentadores". In many situations, both sentences were correct but we identified as better the sentences that utilized a more scientific language, more appropriate for a publication, e.g., "nivel de escolaridade" instead of just "escolaridade". In another of such cases, we chose the term "longevos" as more appropriate than "mais velhos" when referring to elderly people. We also found errors due to nominal concordance with the number, such "dividido" when related to plural nouns, when it should have been "divididos".

Some mistakes were very subtle, such as the translation shown below which includes the verb "apresentaram" twice in the same sentence. Further, in the translated sentence, it is not clear whether the first instance of the verb "apresentaram" (present) refers just to the second or both subjects, while this information is clear in the reference translation, i.e. that it should refer just to "casos". However, this ambiguity is also present in the original English sentence.

> ***Tumors larger than 2cm and cases that presented angiolymphatic invasion had...***
> – Source
>
> *Tumores maiores do que 2cm e casos com invasão angiolinfática apresentaram...*
> – Reference translation
>
> *Tumores maiores que 2cm e casos que apresentaram invasão angiolinfática apresentaram...*
> – Translation

Another subtle mistake that we found relates to the meaning of the sentence which changed in the translation. In the first sentence below, the subject of the sentence is unknown, while in the second one it is clear that the elderly people are the ones who provide the information.

> *Identificar e hierarquizar as dificuldades referidas no desempenho das atividades de vida diária de idosos.*
> – Sentence 1
>
> *Identificar e hierarquizar as dificuldades relatadas pelos idosos na realização das atividades de vida diária.*
> – Sentence 2

Spanish. The overall quality of the Spanish translations was uneven across all four systems submitted to the challenge. BSC and Talp_upc MT systems had a very good performance when compared to the reference translation, with being BSC the best of the four. UCAM MT's system had a reasonable performance but MT-UOC-UPF was the most irregular.

Sentence structure and word order have shown very good results in all systems for short sentences as shown in the following example.

> ***Isotretinoin is still the best treatment for severe nodulocystic acne.***
> – Source
>
> *la isotretinoína todavía es el mejor tratamiento para el acné noduloquístico severo.*
> – Reference translation
>
> *La isotretinoína sigue siendo el mejor tratamiento para el acné noduloquístico severo.*
> – Translation C

However this was not the case of all sentences, some of which followed English word order, resulting in grammatical correct but unnatural sentences in the target language. Other frequent problems include the handling of acronyms (e.g. *EDs*) and additional information included in the reference translation that was not present in the source, as shown in the example below. (cf. *N = 480*)

> ***Ten Eds will be randomly assigned to the intervention group and 10 to the***

control group.
– Source

Se asignará de forma aleatoria 10 SU ***(N = 480)*** *al grupo de intervención y 10 SU* ***(N = 480)*** *al grupo de control.*
– Reference translation

Diez EDs se asignarán aleatoriamente al grupo de intervención y 10 al grupo de control.
– Translation D

Erroneous word order translation for technical terms has been observed resulting in mistranslation of the English source (e.g. *FE-IV*) sentence as shown bellow.

Additionally, system A has translated *fixed-effects instrumental-variable* as *efectos fijos variable instrumental*, that not only is a mistranslation of this technical term, but also changes the overall meaning of the sentence.

Fixed-effects instrumental-variable (FE-IV) pseudo-panel estimation from three rounds of the Mexican National Health and Nutrition Survey (2000, 2006 and 2012).
– Source

Estimación de pseudopanel de variables instrumentales de efectos fijos (FE-IV) en tres rondas de la Encuesta Nacional de Salud y Nutrición de México (2000, 2006 y 2012).
– Reference translation

Los efectos fijos variable instrumental (FE-VI) se estimaron en tres rondas de la Encuesta Nacional de Salud y Nutrición de México (2000, 2006 y 2012).
– Translation A

Subject-verb agreement mistakes have been observed in some MT translations, such as the one that follows.

Each group will enroll 480 patients, and the outcomes will be compared between groups.
– Source

Cada grupo incluirán 480 pacientes y los resultados serán comparados entre grupos.
– Translation B

Other issues found, more common in longer sentences, are missing information in the translation or wrongly parsed and separated terms, especially if the source sentence also suffers from the same problem.

For the 5-year time horizon, the incrementalcost per patient with mirabegron 50 mg versustolterodine was 195.52 and 157.42, from theNational Health System (NHS) and societal perspectivesrespectively, with a gain of 0.0127 QALY withmirabegron.
– Source

Para el horizonte temporal de 5 años, el incremento por paciente con mirabegron 50 mg versustolterodina fue 195,52 y 157,42 , del Sistema Nacional de Salud (SNS) y de la perspectiva social respectivamente, con una ganancia de 0,0127 AVAC con mirabegron.
– Translation D

8 Conclusions

We presented the 2019 edition of the WMT shared task for biomedical machine translation. Participants were challenged to provide automatic translations for medical texts from the literature in ten language pairs as well as for terminology content from English to Spanish. We prepared three baseline systems based on neural toolkits and received 59 runs from 11 teams. Overall, submissions were received for all test sets that were offered. Some of the results obtained by the participants could outperform the scores from previous editions of the shared task and some submissions were judged better than the reference translations created by the authors of the papers in the test set. We also identified some limitations of this shared task, such as issues with the quality of the test sets that we plan to improve in the next edition of the task. Other planned improvements include manual evaluation

of the submission based on direct assessment as opposed to the current pairwise comparison of two sentences.

Acknowledgments

We would like to thank all participants in the challenges, and especially those who supported us for the manual evaluation, including Melana Uceda (es/en). We also would like to thank the participants Aihu Zhang (team OOM), Antoni Oliver Gonzalez (team MT-UOC-UPF), Fabien Cromières (team KU), Sadaf Abdul-Rauf (team NRPU) and Stefania Duma (team UHH-DS) for providing summaries about their systems, which we included in the manuscript. MK and FS acknowledge support from the *encargo de gestión* SEAD-BSC-CNS of Plan for the Advancement of Language Technology (Plan TL) and the Interreg Sudoe ICTUSnet project.

References

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

Sadaf Abdul-Rauf, Holger Schwenk, Patrik Lambert, and Mohammad Nawaz. 2016. Empirical use of information retrieval to build synthetic data for smt domain adaptation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):745–754.

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.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio.

Mirela-Stefania Duma and Wolfgang Menzel. Translation of Biomedical Documents with Focus on Spanish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*.

Mirela-Stefania Duma and Wolfgang Menzel. 2016a. Data Selection for IT Texts using Paragraph Vector. In *Proceedings of the First Conference on Machine Translation*, pages 428–434, Berlin, Germany.

Mirela-Stefania Duma and Wolfgang Menzel. 2016b. Paragraph Vector for Data Selection in Statistical Machine Translation. In *Proceedings of the 13th Conference on Natural Language Processing KONVENS 2016*, pages 84–89, Bochum, Germany.

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.

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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 67–72, Vancouver, Canada.

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 ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In

Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14, pages II–1188–II–1196. JMLR.org.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

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)*. European Language Resource Association.

Mariana Neves. 2017. A parallel collection of clinical trials in portuguese and english. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 36–40. Association for Computational Linguistics.

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 2016)*, Paris, France. European Language Resources Association (ELRA).

Noor-e-Hira, Sadaf Abdul Rauf, Kiran Kiani, Ammara Zafar, and Raheel Nawaz. 2019. Translation of Medical texts using Transfer Learning and Selective Data Training: NRPU-FJ Participation in WMT19. In *Proceedings of the Fourth Conference on Machine Translation*. Association for Computational Linguistics.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, and Zdeňka Urešová. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine*, 61(3):165 – 185. Text Mining and Information Analysis of Health Documents.

Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Ales Tamchyna, Andy Way, and Josef van Genabith. 2015. Domain adaptation of statistical machine translation with domain-focused web crawling. *Language resources and evaluation*, 49(1):147–193. 26120290[pmid].

Casimiro Pio Carrino, Bardia Rafieian, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. Terminology-aware segmentation and domain feature data enriching strategy for the WMT19 Biomedical Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2019. UCAM Biomedical translation at WMT19: Transfer learning multi-domain ensembles. In *Proceedings of the Fourth Conference on Machine Translation*. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Felipe Soares and Martin Krallinger. 2019. BSC participation in the WMT biomedical task. In *Proceedings of the Fourth Conference on Machine Translation*. Association for Computational Linguistics.

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

Felipe Soares, Gabrielli Harumi Yamashita, and Michel Jose Anzanello. 2018b. A parallel corpus of theses and dissertations abstracts. In *International Conference on Computational Processing of the Portuguese Language*, pages 345–352. Springer.

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).

Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. *JMIR Public Health Surveill*, 5(1):e11171.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the*

Americas (Volume 1: Research Papers), pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

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 *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimón, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Mingxuan Wang, Li Gong, Wenhuan Zhu, Jun Xie, and Chao Bian. 2018. Tencent neural machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 522–527, Belgium, Brussels. Association for Computational Linguistics.

Peng Wei, Liu Jianfeng, Li Liangyou, and Liu Qun. 2019. Huawei's NMT systems for the WMT 2019 biomedical translation task. In *Proceedings of the Fourth Conference on Machine Translation*. Association for Computational Linguistics.

Cuijun Wu, Fei Xia, Louise Deleger, and Imre Solti. 2011. Statistical machine translation for biomedical text: are we there yet? *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2011:1290–1299. 22195190[pmid].

Weijia Xu and Marine Carpuat. 2018. The university of Maryland's Chinese-English neural machine translation systems at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 535–540, Belgium, Brussels. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).