Findings of the WMT 2018 Biomedical Translation Shared Task: Evaluation on Medline test sets

Mariana Neves
German Federal Institute for Risk Assessment (BfR), Germany

Antonio Jimeno Yepes
IBM Research, Australia

Aurélie Névéol
LIMSI, CNRS, Uni. Paris Saclay, France

Cristian Grozea
Fraunhofer Institute FOKUS, Germany

Amy Siu
Beuth University of Applied Sciences, Germany

Madeleine Kittner
Humboldt-Universität zu Berlin, Germany

Karin Verspoor
University of Melbourne, Australia

Abstract

Machine translation enables the automatic translation of textual documents between languages and can facilitate access to information only available in a given language for non-speakers of this language, e.g. research results presented in scientific publications. In this paper, we provide an overview of the Biomedical Translation shared task in the Workshop on Machine Translation (WMT) 2018, which specifically examined the performance of machine translation systems for biomedical texts. This year, we provided test sets of scientific publications from two sources (EDP and Medline) and for six language pairs (English with each of Chinese, French, German, Portuguese, Romanian and Spanish). We describe the development of the various test sets, the submissions that we received and the evaluations that we carried out. We obtained a total of 39 runs from six teams and some of this year's BLEU scores were somewhat higher that last year's, especially for teams that made use of biomedical resources or state-of-the-art MT algorithms (e.g. Transformer). Finally, our manual evaluation scored automatic translations higher than the reference translations for German and Spanish.

1 Introduction

Automatic translation of documents from one language to another facilitates broader information access for resources only available in a particular language. Even in the scientific literature, in which most important articles are published only in English, an increasing number of researchers support citing articles published in other languages for the sake of not missing important research or to avoid carrying out duplicate experiments (Lazarev and Nazarovets, 2018). Recent discussions on this topic in the journal *Nature* have appealed for translation of the best Chinese papers (Tao et al., 2018) and the development of automatic tools for the automatic translation of publications (Prieto, 2018).

Therefore, biomedicine is a domain for which suitable parallel corpora, official evaluation test sets and machine translation (MT) systems are in high demand. There is active development of parallel corpora in this domain (see the recent survey in (Névéol et al., 2018)). In this year alone, three new corpora have been published in a single conference: a compilation of full texts from the Scielo database for English, Portuguese, and Spanish (Soares et al., 2018), medical documents and glossaries for Spanish/English (Villegas et al., 2018) and a biomedical corpus for Romanian (Mitrofan and Tufis, 2018). However, in spite of the growing number of parallel corpora and the many open source tools for MT (e.g., Moses (Koehn et al., 2007), OpenNMT (Klein et al., 2017) and Marian (Junczys-Dowmunt et al., 2018)), there is still no ready-to-use tool for automatic translation of biomedical publications for any language pair.

With the aim of fostering advances in this field, we organized the third edition of the Biomedical Translation Task in the Conference for Machine Translation (WMT).[1] It builds on the two previous editions (Bojar et al., 2016; Jimeno Yepes et al., 2017) by offering test sets from Medline for six

[1] http://www.statmt.org/wmt18/biomedical-translation-task.html

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 328–343

https://doi.org/10.18653/v1/W18-6403

language pairs and from EDP for one language pair, as detailed below:

- Chinese-English (zh/en); Eng.-Chinese (en/zh)
- French-English (fr/en); Eng.-French (en/fr)
- German-English (de/en); Eng.-German (en/de)
- Portuguese-English (pt/en); Eng.-Port. (en/pt)
- English-Romanian (en/ro)
- Spanish-English (es/en); Eng.-Span. (en/es)

Most test sets were derived from scientific abstracts from Medline which were available in both languages. Except for Romanian, we addressed translation in both directions for all language pairs. This was not possible for Romanian due to the low number (less than 50) of parallel abstracts which are available in Medline. For the first time, we have an Asian language, specifically Chinese.

In this paper, we describe details of the challenge. Section 2 presents the construction and quality analysis of the test sets, followed by the details on the six participating teams in Section 3. Section 4 presents the results for both automatic and manual evaluation that we carried out, as well as some additional evaluations which are new this year. Finally, we provide a comprehensive discussion of the results and quality of the translations in Section 5.

2 Test sets

Test sets were obtained from Medline and EDP. In these sources, text for both languages is readily available from the authors of the publications.

EDP. This year's test set was derived from last year's processing of publications. We kept one extra test set for this year's challenge. It can be noted that the sentence segmentation offered for the EDP corpus this year was performed manually. More details can be found in the description of the challenge in 2017 (Jimeno Yepes et al., 2017).

MEDLINE. We constructed the various Medline test sets following a similar strategy carried out for the Scielo corpus (Neves et al., 2016). We started by downloading MEDLINE 2018 and retrieving those entries whose abstract was available for more than one language, usually English was one of the languages. Such abstracts are identified by the XML tag *OtherAbstract* and its attribute *Language*. We only considered the abstract of the publications since the titles were frequently only available in one language. We randomly selected a subset of the abstracts for the six language pairs under consideration and for which we have native speakers of the foreign languages.

The text of the abstracts were extracted from the XML files and 120 abstracts were randomly selected, excepted for Romanian whose total of parallel documents in Medline was less than 50. The number 120 accounts for possible errors in the preprocessing of the abstract in order to have a final test set of 100 abstracts to be split into the two translation directions. The documents were automatically split using the Stanford CoreNLP tool and the respective available models for each languages, i.e., Chinese, French, German and Spanish (Manning et al., 2014).[2] Since for Portuguese and Romanian no models are available in the Stanford CoreNLP tools, we used models for other similar Roman languages (Spanish for Portuguese and French for Romanian). The sentences were then automatically aligned using the GMA tool for which we provided a list of stopwords for each language.[3] After a short analysis of the alignment of the Chinese/English abstracts, and given the bad alignments that we obtained, we carried out a new automatic alignment using the Champollion tool (Ma, 2006).[4] The resulting aligned sentences were then manually checked for assessing their quality.

2.1 Manual evaluation of the automatic alignment

After compiling the Medline test sets, we manually checked the totality of the abstracts to assess the quality of the automatic alignment (cf. results shown in Table 2). We utilized a modified version of the Quality Checking task of our installation of the Appraise tool (Federmann, 2010, 2018) and one native speaker of each non-English language carried out the validation (cf. Figure 1). The only exception were the Chinese abstracts which were manually checked without the use of the Appraise tool. For each language pair, we checked the totality of the abstracts for both translation directions, e.g., en/de and de/en, which was later randomly

[2] https://stanfordnlp.github.io/CoreNLP/

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

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

Test sets		de/en	fr/en	pt/en	es/en	en/de	en/fr	en/pt	en/es	en/ro	en/zh	zh/en
EDP	# documents		86				83					
	# sentences		879/880				823/821					
Medline	# documents	48	49	50	50	48	49	50	50	40	50	49
	# sentences	342/337	318/328	283/286	286/300	352/378	279/281	332/318	299/263	301/293	311/307	279/311

Table 1: Overview of the test sets. We present the number of documents and sentences in each test set. The number of sentences might be different for the two languages in a test set.

split into two test sets. The only exception was the Romanian test set. Due to its small size, we only built one test set for one translation direction (en/ro).

The number of completely unaligned sentences was rather uniform across the various language pairs and usually less than 5%, with the exception of Spanish (more than 8%) and German (more that 15%). All other partial alignments (Overlap, Source>Target and Target>Source) had a contribution of less than 10%. For three languages, at least 80% of the sentences were correctly aligned, while for Spanish and German only 70% and 65% of the sentences were correctly aligned. The lower quality of these two test sets could certainly affect the calculation of the BLEU score and we will address this problem later in Section 4.2.3.

During the manual validation, we detected problems in the parallel abstracts. For instance, four abstracts (PMIDs 24616752, 25767637, 26941877 and 24294348) in the German test set had to be excluded because they were wrongly tagged in Medline as being in German, while they were written in Italian. The same occurred in the French test sets, were two abstracts (PMIDs 23396711 and 24883131) were also in Italian. This suggests that the use of automatic language identification tools could be useful to validate the language metadata retrieved from MEDLINE.

3 Participating teams and systems

We received submissions from six teams, as summarized in Table 3. The teams came from research and academic institutions of four countries (Brazil, Germany, Spain and USA) and from three continents. An overview of the teams and their systems is provided below.

FOKUS (Fraunhofer Institute FOKUS, Germany). The FOKUS team participated with a system based on neural machine translation (NMT) based on the implementation of the Transformer architecture (Kaiser et al., 2017; Vaswani et al., 2017) for MT (Grozea, 2018). The NMT system made use of biomedical and news corpora for either training or validation (tuning). In addition to this, and in order to automatically select the highest fidelity translation, they developed heuristics based on a dictionary and on stemming. Further, they performed diacritics normalization in order to account for recent ortographic changes in the Romanian language.

Hunter MT (Hunter College, USA). The Hunter team (Khan et al., 2018) used different transfer learning methods and trained different in-domain biomedical data sets one after another. Their system was set up using parameters of previous training as the initialization of the following training. A News based model was used as pre-training.

LMU (Ludwig Maximilian University of Munich, Germany). The LMU team implementated various neural network models and trained and tuned the models on parallel biomedical data (Huck et al., 2018). They experimented with implementations of the Transformer architecture (Sockeye implementation) and the encoder-decoder models (Nematus toolkit). The authors highlight that the word segmentation used on the German language for both translation directions were responsible for the good performance of the system in the human evaluation.

TFG TALP UPC (Technical University of Catalunya, Spain). For their system that provides translations into English, the TGF TALP UPC team participated with a Transformer architecture (Kaiser et al., 2017; Vaswani et al., 2017) using both single-language and multi-source systems (Tubay and Costa-Jussà, 2018). The systems were trained on the Scielo and Medline titles made available by the shared task in the last years. The multi-source systems utilized a concatenation of training data from es/en, fr/en and pt/en.

UFRGS (Universidade Federal do Rio Grande do Sul, Brazil). The UFRGS team participated with two runs based either on Moses (Koehn et al.,

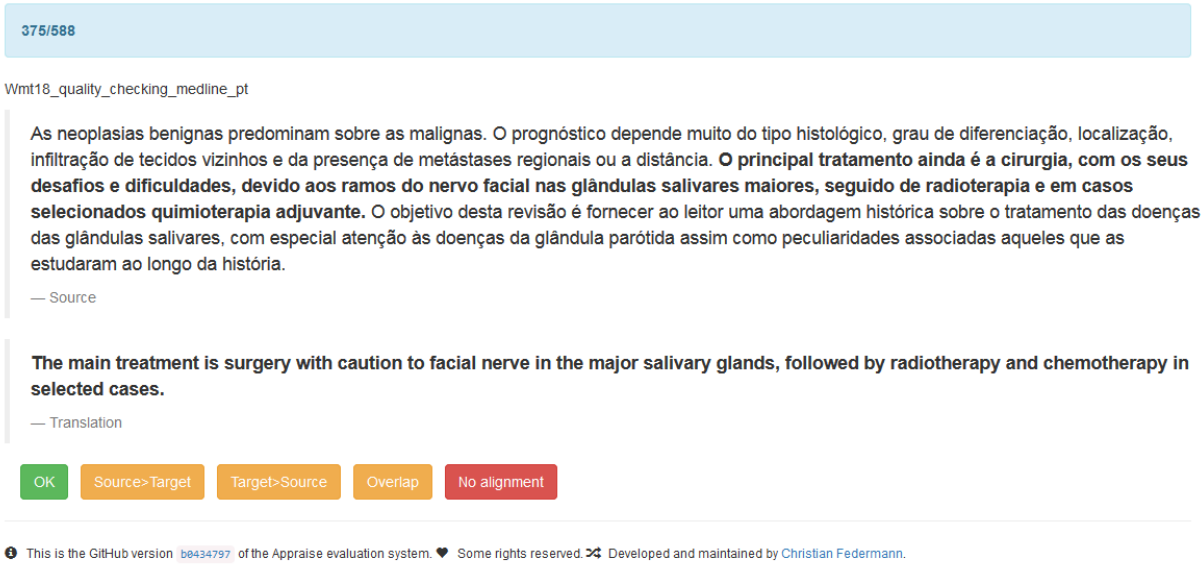

Figure 1: Screen-shot of Appraise during manual validation of the dataset for Portuguese.

Test sets	No alignment	OK	Overlap	Source > Target	Target > Source	Total
en/de, de/en	104 (15.38%)	437 (64.64%)	23 (3.40%)	60 (8.88%)	52 (7.69%)	676
en/es, es/en	46 (8.30%)	388 (70.04%)	38 (6.86%)	44 (7.94%)	38 (6.86%)	554
en/fr, fr/en	20 (3.36%)	528 (88.59%)	6 (1.01%)	20 (3.36%)	22 (3.69%)	596
en/pt, pt/en	11 (1.87%)	490 (83.33%)	9 (1.53%)	41 (6.97%)	37 (6.29%)	588
en/ro	7 (2.39%)	260 (88.74%)	3 (1.02%)	12 (4.10%)	11 (3.75%)	293
en/zh, zh/en	19 (3.26%)	528 (90.72%)	4 (0.69%)	18 (3.09%)	13 (2.23%)	582

Table 2: Manual validation of the automatic alignment sentences for the Medline test sets. Values are shown in absolute and percentage numbers. The test include the abstracts for both languages directions, with the exception of the Romanian language. The total column represents the totality of the aligned sentences

Team ID	Institution
FOKUS	Fraunhofer Institute FOKUS (Germany)
Hunter MT	Hunter College (USA)
LMU	Ludwig Maximilian University of Munich (Germany)
TFG TALP UPC	Technical University of Catalunya (Spain)
UFRGS	Universidade Federal do Rio Grande do Sul (Brazil)
UHH-DS	University of Hamburg (Germany)

Table 3: List of the participating teams.

2007) or OpenNMT (Klein et al., 2017) systems (Soares and Becker, 2018). Training data was prepared by concatenating several in-domain and out-of-domain resources. The in-domain corpora included scientific articles (full texts) from Scielo, the UFAL medical corpus, the EMEA corpus and Brazilian theses and dissertations. Due to possible overlap with the test sets from Medline, the team applied some procedures to automatically exclude some publications from the Scielo training data. Terminological resources such as the Unified Medical Language System (UMLS) (Bodenreider, 2004) were used as well.

UHH-DS (University of Hamburg, Germany). The UHH-DS team utilized Moses (Koehn et al., 2007) trained on a variety of in-domain and general domain corpora (Duma and Menzel, 2018). The main feature of their system was the development of an unsupervised method to automatically under-sample sentences from the general domain collection that were better suited for the biomedical domain. Their under-sampling algorithm can be applied either on the source or target side of the corpora, as well as on both sides.

4 Evaluation

In this section we describe the various submissions that we obtained and present the results that these achieved based on both automatic and manual valuation.

4.1 Submissions

In total, we received 39 submissions from the six teams, as summarized in Table 4. Unfortunately, we received no submissions for Chinese (neither zh/en nor en/zh) and no submissions for the French EDP test set (fr/en).

FOKUS. The FOKUS team submitted two runs in which one (run1) was trained on a biomedical corpus and validated on news corpora while the second one (run2, primary run) is an ensemble of various NMT systems and uses the heuristics they defined for selecting the best translation.

Hunter. The Hunter's team submitted two runs for en/fr for each of the Medline and EDP test sets. In these runs, they considered NMT based ensembles and trained on various in-domain and out-of-the-domain corpora. However, differences between the runs are unclear.

LMU. The three en/de submissions from the LMU team were the following: a right-to-left re-ranked Transformer (run1, primary run), a Transformer ensemble without re-ranking (run2) and the encoder-decoder built with Nematus (run3). The only submission for de/en was a Transformer without ensemble.

TFG TALP UPC. Each two submissions for language pairs es/en, fr/en and pt/en utilized either multi-source (run1, primary run) or the single-source (run2) training.

UFRGS. The two submissions from the UFRGS teams seem to have differed only on the MT tool that they used, i.e., either OpenNMT (run1, primary run) or Moses (run2).

UHH-DS. The three submissions for each of the language pairs (en/es, en/pt, en/ro, es/en and pt/en) differed on whether the under-sampling algorithm was applied only on the English side (run1), on the non-English side (run2) or on both sides (run3, primary run).

4.2 Automatic evaluation

Here we provide the results for the automatic evaluation and rank the systems regarding the resulting scores. We computed BLEU scores at the sentence level using the script `mteval-v14.pl` from the Moses distribution.[5] For all test sets and language pairs, we compare the submissions (automatic translations) to the respective reference one.

4.2.1 Automatic evaluation: EDP test sets

The BLEU scores for the EDP test set are presented in Table 5. Given that we received only two submissions from a single team, we could not perform comparison between teams. We ranked the two submissions as follows:

- en/fr: Hunter (run 1) $<$ Hunter (run 2).

Run2 obtained a slightly higher score than run1, however, reasons for this improvement are unknown.

4.2.2 Automatic evaluation: Medline test sets

This year, we calculated BLEU scores based on the totality of the sentences (including the ones with incorrect alignments) as well as based only

[5] http://www.statmt.org/moses/?n=Moses.SupportTools

Teams	de/en	en/de	en/es	en/fr	en/pt	en/ro	es/en	fr/en	pt/en	Total
FOKUS						M2				2
Hunter				E2M2						4
LMU	M	M3								4
TFG TALP UPC							M2	M2	M2	6
UFRGS			M2		M2		M2		M2	8
UHH-DS			M3		M3	M3	M3		M3	15
Total	1	3	5	4	5	5	7	2	7	39

Table 4: Overview of submissions for each language pair and test set: [M]edline and [E]DP. The number next to the letter indicates the number of runs that the team submitted for the corresponding test set (if larger than one).

Team	Runs	en/fr
Hunter	run1	22.20
	run2	23.24*

Table 5: BLEU scores for the EDP en/fr dataset. * indicates the primary run as informed by the participants.

on the sentences which were perfectly aligned (cf. Section 2).

BLEU scores for the Medline test set are presented in Table 6. For some language pairs, i.e., de/en, en/de, en/fr and fr/en, we could not compare results between various teams since we received submissions only from one team. Moreover, we only received one submission from one team for de/en. Therefore, no further comparison was possible for this language pair. We ranked the various teams and submissions, for those languages for which we received more than one submission, as follows:

- en/de: LMU (run3) < LMU (runs 1,2);
- en/es: UHH-DS (runs 1,2,3) < UFRGS (runs 1,2);
- en/fr: Hunter (runs 1,2);
- en/pt: UHH-DS (runs 1,2,3) < UFRGS (runs 1,2);
- en/ro: UHH-DS (runs 1,2,3) < FOKUS (run 1) < FOKUS (run 2);
- es/en: UHH-DS (run 2) < UHH-DS (runs 1,3) < TGF TALP UPC (runs 1,2) < UFRGS (runs 1,2);
- fr/en: TGF TALP UPC (run 2) < TGF TALP UPC (run 1);
- pt/en: TGF TALP UPC (run 2) < TGF TALP UPC (run 1) < UHH-DS (runs 1,2,3) < UFRGS (runs 1,2).

In the following we provide a short summary of the results with regard to the method or resources that have been used.

de/en. The run based on the Transformer architecture from the LMU team obtained a reasonable BLEU score. However, we could not compare this to any other submission.

en/de. There was little difference in the BLEU score between the two first submissions, both based on the Transformer architecture, but both did seem to be superior to the third run based on the encoder-decoder model.

en/es. The best results for en/es were obtained by the UFRGS team when using the Moses system (run2) instead of neural MT (run1), as expected by the team. However, the difference between both submissions is not significant. We observed no significant difference between the three submissions from the UHH-DS team. However, all of them yield much lower BLEU scores than the submissions by the UFRGS team.

en/fr. The submissions from the Hunter team obtained very similar scores for the Medline test sets. Details on each run is unclear but these differences seem to have brought significant improvement on the scores only on the EDP test set (cf. Table 5).

en/pt. Both submissions from the UFRGS team obtained the highest BLEU scores, which again and similar to the results obtained for en/es, did not confirm the superiority of neural MT. The three runs from the UHH-DS team were closer to the ones from the UFRGS team (in comparison to the ones for en/es), but still rather inferior. This time, run1 (under-sampling based on the English side) did perform a little better than the other two runs, specially regarding run2 (under-sampling based on the non-English side).

Teams	Runs	de/en	en/de	en/es	en/fr	en/pt	en/ro	es/en	fr/en	pt/en
FOKUS	run1						16.97			
	run2						18.10*			
Hunter MT	run1				23.41					
	run2				23.24*					
LMU	run1	23.93*	18.81*							
	run2		18.75							
	run3		17.16							
TFG TALP UPC	run1							40.49*	25.78*	39.49*
	run2							39.06	19.42	38.54
UFRGS	run1			39.62*		39.43*		43.31*		42.58*
	run2			39.77		39.43		43.41		42.58
UHH-DS	run1			31.32		34.92	14.60	36.16		41.84
	run2			31.05		34.19	14.39	35.17		41.80
	run3			31.33*		34.49*	14.07*	36.05*		41.79*

Table 6: Results for the Medline dataset. * indicates the primary run as informed by the participants.

en/ro. The run2 of the FOKUS team, which consisted on an ensemble of various NMT systems and used heuristics for selecting the best translation, obtained the highest BLEU score. The two submissions from UHH-DS reached BLEU scores which were slightly below results from the FOKUS team. We observed no significant difference between the three runs. However, similar to en/pt, under-sampling based on the English side (run1) seems to perform slightly better than under-sampling based on both sides (run3).

es/en. Once again the statistical MT system (Moses) from the UFRGS team obtained a slightly higher score than their neural MT system (OpenNMT). Indeed, the BLEU scores obtained by run2 of the team was the highest one among all submissions to the shared task for all language pairs. The following two best scores belonged to the Transformer-based MT systems from the TGF TALP UPC team. Even though a more recent and currently state-of-the-art method (Transformer) was used by team TGF TALP UPC, the better results obtained by UFRGS were probably due to the larger training collection that they used. The model trained on various sources obtained a slightly better score. The three runs from the UHH-DS team were rather inferior than the ones from the two other teams. A significant difference was only observed for run2 (under-sampling based on the non-English side) which achieved lower BLEU scores for this language pair.

fr/en. The only submissions for fr/en belonged to the Transformer-based MT systems from the TGF TALP UPC team. This time, the improvement of the multi-source model over the single model was very significant. However, the highest score was rather low in comparison to the other submissions of the team.

pt/en. There was no difference in the two submissions from the UFRGS team, both obtained the highest BLEU scores. The three runs from the UHH-DS team obtained the second best results but we observed no significant difference between the three of them. Finally, the two lowest scores were obtained by the Transformer-based MT systems from the TGF TALP UPC team. Similar to results for es/en and fr/en, the system trained on various sources obtained a little improvement over the single models. Also similar to the es/en results, the higher performance from UFRGS was probably due to the use of more resources for training the MT systems.

4.2.3 Evaluation for sentences with good alignment

This year, we also calculated additional BLEU scores when considering only the sentences whose alignments we manually classified as being correct. Correctly aligned means that sentences in both languages contained exactly the same information and neither of the sentences contained more information than the other (cf. Section 2).

Results for this subset of the test set are presented in Table 7. For most teams, improvements were significant, ranging from two to four BLEU points, but was up to one point for en2ro. The overall order of the results mostly remained the same.

4.2.4 Evaluation for Romanian after diacritics normalization

In the particular case of the Romanian language, there were fairly recent changes in the ortho-

Teams	Runs	de/en	en/de	en/es	en/fr	en/pt	en/ro	es/en	fr/en	pt/en
FOKUS	run1						17.84			
	run2						19.11*			
Hunter	run1				24.66					
	run2				24.76*					
LMU	run1	28.84*	24.30*							
	run2		23.88							
	run3		21.84							
TFG TALP UPC	run1							42.91*	27.10*	42.55*
	run2							41.26	20.20	41.56
UFRGS	run1			44.50*		43.14*		46.92*		46.01*
	run2			44.50		43.14		46.92		46.01
UHH-DS	run1			34.77		37.24	15.85	38.45		44.28
	run2			34.70		36.76	15.62	37.17		44.32
	run3			35.08*		36.91*	15.28*	38.18*		44.27*

Table 7: Results for the Medline dataset using OK aligned sentences. * indicates the primary run as informed by the participants.

graphic recommendation with respect to diacritics notation[6]: comma-below ș and Ș should be used instead of cedilla-below ş and Ş; in the same way, the comma-below ț and Ț should be used instead of the cedilla-below ţ and Ţ, according to a 2003 communicate from the "Iorgu Iordan" Institute of Linguistics of the Romanian Academy.

While the two comma-below and cedilla-below variants of those letters are hardly distinguishable to a human reader, they have different unicode codes and thus replacing one with another in a word makes it a completely different word, for an automated method. Having the "wrong" word affects all n-grams containing that word for the BLEU scoring.

In order to achieve more quality in the translation assessment, we normalized all diacritics both in gold standard and in the submissions for Romanian. Results for the Medline en/ro test set are shown in Table 8, based on all sentences (en/ro) and only based on correctly aligned sentences (en/ro-OK). To this end, we wrote and used a simple sed-based script which brings the Romanian diacritics to the latest standard[7].

4.3 Manual evaluation

We performed manual evaluation of the primary runs (as identified by the participants) for each team and each language pair. The primary runs are compared to the reference translation and to each other, if more than one submission (from distinct teams) is available for the language pair. We computed pairwise combinations of translations either between two automated systems, or one automated system and the reference translation. The human validators were native speakers of the languages and were either members of the participating teams or colleagues from the research community. These are primary runs from each team:

- FOKUS: Medline en/ro run2;
- Hunter: Medline en/fr run2, EDP en/fr run2;
- LMU: Medline de/en run1, Medline en/de run1;
- TGF TALP UPC: Medline es/en run1, Medline fr/en run1, Medline pt/en run1;
- UFRGS: Medline en/es run1, Medline en/pt run1, Medline es/en run1, Medline pt/en run1;
- UHH-DS: Medline en/es run3, Medline en/pt run3, Medline en/ro run3, Medline es/en run3, Medline pt/en run3.

The validation task was carried out using the 3-way ranking task in our installation of the Appraise tool (Federmann, 2010).[8] For each pairwise

Teams	Runs	en/ro	en/ro-OK
FOKUS	run1	22.17	22.98
	run2	23.42*	24.22*
UHH-DS	run1	15.40	15.95
	run2	15.09	15.69
	run3	14.77	15.44

Table 8: Results for the Medline en2ro test set after normalization of diacritics. * indicates the primary run as informed by the participants.

[6]For reference, the evolving standards for Romanian are discussed here http://kitblog.com/2008/10/romanian_diacritic_marks.html.

[7]The script is freely available at http://www.brainsignals.de/fixrodia.sh

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

comparison, we checked a total of 100 randomly-chosen sentence pairs. The validation consisted of reading the two sentences (A and B), i.e., translations from two systems or from the reference, and choosing one of the options below:

- A<B: when the quality of translation B was higher than A.
- A=B: when both translations had similar quality.
- A>B: when the quality of translation A was higher than B.
- Flag error: when the translation did not seem to be derived from the same input sentence. This is usually related to errors in corpus alignment.

We present the results for the manual evaluation of the Medline test sets in Table 9. Based on the number of times that a translation was validated as being better than another, we ranked the systems for each language as listed below:

- de/en: LMU = reference
- en/de: reference < LMU
- en/es: reference, UHH-DS < UFRGS
- en/fr: HunterNMT < reference
- en/pt: UHH-DS < UFRGS < reference;
- en/ro: UHH-DS < FOKUS < reference
- es/en: UHH-DS < UFRGS < TGF TALP UPC < reference
- fr/en: TGF TALP UPC < reference
- pt/en: UHH-DS < UFRGS < TGF TALP UPC < reference

Even though the LMU runs obtained one of the lowest BLEU scores (all of them less than 20 points), the primary run did score equally well or even better than the reference translation in the manual evaluation. The reason are misaligned sentences in the German reference. Automatic German translations on the other hand are most often correct in translation and alignment of content. Indeed, the quality of the German dataset was one of the lowest (cf. Section 2.1).

We present the results for the manual evaluation of the EDP test sets in Table 10. Based on the number of times that the submission was validated as being better than the reference translation, we ranked the two translations as follow:

- en/fr: Hunter < reference.

5 Discussion

In this section we present insight from the automatic and manual validations as well as on the quality of the translations.

5.1 Differences between manual and automatic evaluations

Similar to previous years, we did not notice any difference while ranking the teams for most language pairs regarding the automatic and manual evaluation of the translations. This year, the only significant difference we noticed was for the English translations, more specifically for es/en and pt/en pairs.

For es/en, the ranking order changed between the teams UFRGS and TGF TALP UPC. While the runs from the UFRGS teams achieved a higher BLEU score (43.31 vs. 40.49), our evaluators found the translations from the TGF TALP UPC team to be considerable better (79 vs. 7).

As for pt/en, the ranking of the teams changed from TGF TAP UPC < UHH-DS < UFRGS (automatic evaluation: 42.55 < 44.27 < 46.01) to UHH-DS < UFRGS < TGF TAP UPC (manual evaluation: 55 vs. 21 to UFRGS, 58 vs. 24 to UHH-DS). While no difference in ranking was observed between teams UHH-DS and UFRGS, in comparison to the automatic evaluation, team TGF TAP UPC moved from being the last ranked in the automatic evaluation to the best ranked one on the manual evaluation.

We can only hypothesize that the better BLEU scores that the UFRGS team obtained were probably due to better translation of particular concepts or due to using the same terms as in the reference translations. However, the TGF TAP UPC team could obtain higher quality of the manual translations using their Transformer architecture. The better performance of the TGF TAP UPC team could also have been due to the test set being included in the their training corpus, i.e. overlaps between Medline and the Scielo databases. While both teams trained on the Scielo corpus,

Languages	Runs (A vs. B)	Total	A>B	A=B	A<B
de/en	LMU vs. reference	75	29	14	32
en/de	LMU vs. reference	76	29	32	15
en/es	UFRGS vs. reference	86	37	23	26
	UFRGS vs. UHH-DS	88	29	37	22
	reference vs. UHH-DS	92	30	33	29
en/fr	Hunter vs. reference	92	14	13	65
en/pt	UFRGS vs. reference	86	6	43	42
	UFRGS vs. UHH-DS	100	32	53	15
	reference vs. UHH-DS	81	46	28	7
en/ro	FOKUS vs. reference	88/81	11/14	19/14	58/53
	FOKUS vs. UHH-DS	100/97	57/55	31/27	12/15
	reference vs. UHH-DS	88/85	80/78	6/6	2/1
es/en	TGF TALP UPC vs. reference	72	26	12	34
	TGF TALP UPC vs. UFRGS	100	51	38	11
	TGF TALP UPC vs. UHH-DS	98	79	12	7
	reference vs. UFRGS	77	50	15	12
	reference vs. UHH-DS	77	54	10	13
	UFRGS vs. UHH-DS	100	45	24	31
fr/en	TGF TALP UPC vs. reference	85	24	19	42
pt/en	TGF TALP UPC vs. reference	89	25	26	38
	TGF TALP UPC vs. UFRGS	100	55	24	21
	TGF TALP UPC vs. UHH-DS	100	58	24	18
	reference vs. UFRGS	87	42	22	23
	reference vs. UHH-DS	87	52	28	7
	UFRGS vs. UHH-DS	100	48	27	25

Table 9: 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. Two evaluators (both participants) carried out the validation of the Romanian dataset and results from both of them are shown (separated by a slash).

Languages	Runs (A vs. B)	Total	A>B	A=B	A<B
en/fr	Hunter vs. reference	91	11	26	54

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

the UFRGS team reported that they tried to remove potential overlaps between Medline and Scielo (Soares and Becker, 2018). Overlaps of Medline and Scielo do not explain the lower BLEU scores obtained by the TGF TAP UPC team.

5.2 Differences across languages

Similar to previous years, comparison of results across languages did not provide any unexpected insight. The languages pairs which obtained higher BLEU scores (above 30 points), i.e. en/es, en/pt, es/en and pt/en, were also the ones for which more training data specific for the biomedical domain is available. Indeed, teams that participated with the same system for different language pairs obtained lower scores for those languages with fewer resources. This is the case of the scores of the TGF TAP UPC team for fr/en (up to 27 points) as opposed to the ones obtained for es/en and pt/en (more than 40 points).

We hope that recently released corpora, e.g. the BioRo corpus for Romanian (Mitrofan and Tufis, 2018), can boost performance of MT systems for these languages. However, more parallel corpora are certainly necessary not only for those languages that scored worst in this challenge, but also for the many other languages that we did not evaluate here. Unfortunately, open-access databases such as Scielo are not available for most languages. Nevertheless, the number of parallel abstracts in Medline are increasing and corpora derived from these are starting to being published, e.g. MeSpEn (Villegas et al., 2018).

5.3 Evolution of the performance in the last years

Compared to previous years, we found an improved BLEU score and improved manual evaluation for languages already considered in previous years, i.e. Portuguese, Spanish and French. This year we have considered Medline abstracts instead of Scielo ones.

However, the results are difficult to be directly compared to previous years given that test sets were from different sources for many of the language pairs. For the EDP test set, which can be considered very similar to last year's one, the Hunter team scored much better than their last participation both in terms of BLEU scores (17.50 vs. 23.24 for en/fr) and in the manual validation (0 to 93 vs. 11 to 54 in the manual validation).

The Medline test sets for es/en, pt/en, fr/en, en/es, en/pt and en/fr can be considered rather similar to the Scielo ones released for these language pairs in the two previous challenges (Bojar et al., 2016; Jimeno Yepes et al., 2017). From values below the 20 points in 2016, results from en/pt jumped to almost 40 in 2017 and over the latter (up to 43.14) this year. A similar increase was observed for en/es that increased more slowly from up to 33 points in 2016, up to 36 points in 2017 and up to 44 this year. On the other hand, not much improvement can be noticed from en/fr in 2016 (up to 22.75) to this year's best score (23.24). These values are also similar to the scores reported on another MEDLINE dataset in 2013 (Jimeno-Yepes and Névéol, 2013).

Regarding translations into English, for es/en, BLEU scores experienced an improvement from 37 to 43 points in the last year. However, the same could not be noticed for pt/en that remained rather constant around 41-43 points.

Finally, during our manual validation, we observed for the first time that the quality of some automatic translations was either equal or better than the reference translation. Two teams scored as good as the reference translation, namely, LMU for de/en and UHH-DS for en/es. Moreover, two teams scored higher than the reference translations, namely, LMU for en/de and UFRGS for en/es.

5.4 Quality of the automatic translations

Here we provide an overview of the quality of the translations and the common errors that we identified during the manual validation.

English: The English translations appeared in general to have improved qualitatively over prior year submissions. While in prior years translations often contained remnants of untranslated terms from the source language mixed into the translation, this problem was noted less often in this year's evaluations. In addition, systems appeared to make more effective use of capitalisation, avoiding translation of acronyms or attempting to translate an acronym semantically via its expansion.

In light of this overall improvement, a better translation is often decided by subtle, more precise choices of English words this year. For instance, an "increasing trend" is more precise as well as the more customary usage than "accentu-

ating trend"; the "dissemination" of knowledge is likewise a better word choice than "diffusion" of knowledge. Similarly, a study objective "to assess" the level of something would be preferred to "to know" the level.

The automated systems maintained higher fidelity with the original texts than reference translations, with the latter often leaving out portions of the original sentence or restructuring information between contiguous sentences. Since the automated systems strive to translate the complete content of a sentence, they were in many cases perceived to be more accurate due to completeness, even where minor usage errors occurred.

An error that was observed regularly for Spanish to English translations in particular was the lack of a subject pronoun or insertion of a gendered pronoun ("He"/"She") at the start of a sentence where a demonstrative pronoun ("It", "This") would be more appropriate. As a pro-drop language, the source Spanish texts often lacked an overt subject; this subject needs to be introduced for the English translation to be fully grammatical but some systems appeared to struggle with this requirement.

Another error observed across different languages was the partial translation of multi-word biomedical terms. As an example, "upper digestive endoscopy" was translated as "high digestive endoscopy," where presumably "digestive endoscopy" was referenced from a biomedical dictionary but the word "high" was decoupled from the multi-word term. Although this error was less prevalent, its occurrences critically reduced the quality and crippled the scientific meaning of the translation.

French: The quality of translations for French seemed quite equivalent to last year, and varied from poor to good. A number of automatically translated sentences carried out the meaning of the original sentence properly, but were assessed as inferior to the reference for stylistic reasons, because they provided a more literal translation that mimicked the structure of the original sentence. Arguably, those sentences could be considered as useful to grasp the meaning of the original sentence. However, translation omissions were noted in long or complex sentences. For example, the phrase *"potential drug-drug and food-drug interactions"* was translated by *"interactions potentielles entre médicaments et médicaments"*, which does not account for food-drug interactions. A couple of recurring errors also observed in previous years are the lack of translation for acronyms and the erroneous choice of pronoun in translation. For example, *"they"* was systematically translated as *"ils"*, even when *"elles"* was the correct translation based on context.

The use of manual segmentation on the EDP corpus resulted in a number of single word "sentences" corresponding to the titles of the sections in structured abstracts, such as "Introduction:" or "Conclusion:". Unsurprisingly, the systems performed well on these isolated segments (except for one occurrence of *"Materials and Methods"* translated by *"Matériaux et Procédés"* instead of the usual *"Matériel et Méthodes"*), which may contribute to explain the number of instances where the automatic translation and manual reference were considered equivalent. It can be noted that dealing with section titles as isolated segments successfully ensured there were no translation errors linked to failure to identify the section words as isolated titles.

German: Interestingly, for 80% of the sentences automatic translation was evaluated equally good or even better than the German reference. This observed result has different reasons. Often the German reference translation is correct but either contains additional information or misses information present in the English source sentence while the automatic translation does not have this error. As previously mentioned this is strongly related to the frequent alignment errors present in the German dataset. In some cases validation was very difficult as both translations were very good but we still tried to differentiate between them. For instance, "thromboembolic complications" was translated to "thromboembolische Ereignisse" (events) in the reference and to "thromboembolische Komplikationen" (complications) in the automatic translation. In this case the evaluator scored LMU>Reference while also LMU=Reference would be possible.

In general, we only observe minor mistakes in automatic translation. Rarely, we find wrongly translated technical terms such as *cerebrum* wrongly translated to *Gebärmutter* (uterus). Often mistakes originate from a slight misuse of terms with the same overall meaning but different application in the medical domain. For instance, *soft tissue repair* was translated to *Weichteilsanierung*

instead of *Weichteilrekonstruktion*, while the latter is the correct medical term. Similarly, *efficiency of medication* should be *Wirksamkeit von Medikamenten* instead of *Effizienz von Medikamenten*. Compared to submissions in 2017, we did not observe problems in syntax or grammar which could have caused misunderstanding the meaning of the sentence. This year, only the LMU team submitted reaults and already in 2017 their system did not have syntax problems.

Portuguese: We observed both minor and major mistakes in the automatic translations to Portuguese. We classified as minor those errors that did not compromise the overall understanding of the sentence and that were limited to orthography or minor grammatical errors. For instance, we found many wrong spaces separating compound words (e.g., *"amarelo- palha"*) and before commas or the final period (e.g., *"desafio médico , éticas"*). Further, translations from one particular system seemed to consistently start sentences with a lower case (e.g., *"manifestações clínicas de neurofibromatose tipo 1 são variáveis"*). Finally, other frequent minor mistakes were missing definite articles, such as in *"existem três variantes do osteocondroma extraesquelético : condromatose sinovial , condroma para-articular"* instead of *"a condromatose sinovial, o condroma para-articular"*.

We classified as major those mistakes that considerably compromised the understanding of the sentence. These were cases of discordance with number and gender for the adjectives, e.g., *"é um desafio médico , éticas e psicossociais"* instead of *"é um desafio médico, ético e psicossocial"*. There were also verbal discordances, such as in *"houve um caso que foram tratados"* instead of *"houve um caso que foi tratado"*. Further, we found many words that were not translated into Portuguese and remained in English, such as in the passages *"sem tratamento tumor-directed", "Forty-seven casos"*. Also some acronyms were not correctly into Portuguese (e.g., *PET/CT* instead of *PET/TC*), but translations from one of the teams seem to have gotten most acronyms, biomedical terms and numbers right. Finally, given the differences in word ordering between English and Portuguese, this error occurred in passages such as *"pacientes queixa* instead of *"queixa dos pacientes"*.

Some translation were exactly like the reference translations, which makes us suspect that those abstracts could be included in the Scielo corpus used for training data by the systems. However, there were just a couple of such cases and these should not compromise overall evaluation. In spite of the above, we also found very good translations even for complex and long sentences and for biomedical terms with multiple modifiers, such as in *"esclerose múltipla secundária progressiva"*.

Spanish: Considering previous years of the challenge, the translations seem to improve and there are fewer issues compared to previous years. On the other hand, the issues we identified are similar to the ones identified in the Portuguese sets. As with the translations from Spanish into English, there were some cases of source words not being translated into Spanish, as in *"el estado fsico motor 20 Meter Shuttle Run Test"* and *"Substance-induced fueron"*.

Both types of methods seem to suffer from gender and number agreement for determiners as in *"La pulsos mejor en las"* and sometimes for verbs in terms of number as in *"Legendre describen el primer modelo"*, which might be misleading. We also found that some systems displayed a preference for starting sentences with lower case letters; however, different from the case in Portuguese, for the manually evaluated cases there no issues with acronyms or spaces between hyphenated words.

Romanian: This year there fewer participants than in the previous year. Especially regrettable is the absence of the top performers from the University of Edinburgh. The only team which participated last year as well is that of the University of Hamburg, which improved this year by using a training dataset subsampling heuristic in their SMT translator, but trailed again the NMT system in this task.

When manually comparing the translations, we have prefered the ones having the better grammar – for example "Diagnosticul precoce și tratamentul infecției **sunt asociate** (...)" was prefered to "Diagnosticul precoce şi tratamentul infecţiei **este asociat**(...)" for correct subject to verb agreement. Disturbing in translations is the occurence of words that have no correspondent in the original, for example "upon patients' arrival in the post anaesthesia care unit" translated as "*asupra* sosirii pacienților în unitatea de îngrijire a **tuberculozei**".

Totally incorrect translations were observed as well, such as "In this observational study, clinical data, vital signs and comfort parameters were collected from surgical patients who arrived in the PACU." being translated into "În acest studiu observațional, date **contractuale**, semne vitale și parametri **portuari** au fost colectate de la pacienți **morali** care au sosit în PACU.". In such a short sentence there are already three incorrectly translated words. A dictionary-based method would have done better, as there is also no ambiguity involved.

An interesting aspect was observed where the typical preprocessing leads to ambiguities. "MAP" (mean arterial pressure) changes to "map" (like in geographical map) and then is translated as such: "MAP and HR" was translated as "**Harta** şi *hr*".

Another interesting and potentially dangerous error is the mistranslation of the time units. In one case "Haemofiltration was continued postoperatively in the ICU for another 48 h" was translated as "Haemofiltrarea a continuat postoperativ în ICU pentru încă 48 de **ani**" thus replacing hours with years.

In some cases the translations were marked equal because they were equally bad. In general, the intelligibility and fidelity of the translation was preferred to the form (grammar, smoothness, naturalness), and only for equal content the better form prevailed.

6 Conclusions

This was the third year we have organized the WMT biomedical shared task and we found that the performance has been increasing constantly. Improvements in results seem to be due to a variety of reasons, including more in-domain training data and the use of additional methods that consider transfer approaches and ensemble combination of methods.

From an evaluation perspective, we find that the results improve when we consider only sentences that were perfectly aligned instead of considering all the automatically aligned sentences. This shows some limitations on the quality of the automatically generated test sets. On the other hand, the comparative performance of the different participating systems remains the same.

For some of the languages considered, there were limitations in the quantity of available parallel abstracts. Recent publications include parallel corpora from the database that were previously used for obtaining our test sets. These new corpora include Medline parallel abstracts (Villegas et al., 2018) and full texts from Scielo (Soares et al., 2018). Therefore, manual translation for building the future test could be considered in the following editions of the challenge.

Finally, future improvements should also address problems reported by the participants regarding the current format of the test sets. In the three years of the challenge, we have used BioC as the format for data exchange, which seemed to cause some difficulties for sentence alignment. We will evaluate available formats for data exchange with the participants or inspired in other shared task in WMT.

Acknowledgments

We would like to thank the participation of all teams in the challenge and the support of selected participants in the manual validation of the translations.

References

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

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 (WMT16) at the Conference of the Association of Computational Linguistics*, pages 131–198.

Mirela-Stefania Duma and Wolfgang Menzel. 2018. Translation of biomedical documents with focus on Spanish-English. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Christian Federmann. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *In LREC*.

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

Cristian Grozea. 2018. Ensemble of translators with automatic selection of the best translation - the submission of FOKUS to the wmt 18 biomedical translation task -. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2018. LMU Munichs neural machine translation systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Antonio Jimeno-Yepes and Aurélie Névéol. 2013. Effect of additional in-domain parallel corpora in biomedical statistical machine translation. In *Proceedings of the Fourth International Workshop on Health Text Mining and Information Analysis*.

Antonio Jimeno Yepes, Aurélie Névéol, 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. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *CoRR*, abs/1706.05137.

Abdul Rafae Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas. 2018. Hunter NMT system for WMT18 biomedical translation task: Transfer learning in neural machine translation. In *Proceedings of the Third Conference on Machine Translation*, pages 1–2, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vladimir S. Lazarev and Serhii A. Nazarovets. 2018. Don't dismiss citations to journals not published in english. *Nature Correspondence*, 556:174.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *In Proceedings of LREC-2006*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Maria Mitrofan and Dan Tufis. 2018. BioRo: The Biomedical Corpus for the Romanian Language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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).

Daniel Prieto. 2018. Make research-paper databases multilingual. *Nature Correspondence*, 560:29.

Felipe Soares and Karin Becker. 2018. UFRGS participation on the wmt biomedical translation shared task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

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).

Juan Tao, Chengzhi Ding, and Yuh-Shan Ho. 2018. Publish translations of the best chinese papers. *Nature Correspondence*, 557:492.

Brian Tubay and Marta R. Costa-Jussà. 2018. Neural machine translation with the Transformer and multi-source romance languages for the biomedical wmt

2018 task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimn, 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).