Findings of the WMT 2017 Biomedical Translation Shared Task

Antonio Jimeno Yepes
IBM Research Australia

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

Mariana Neves
HPI Uni. Potsdam, BfR, Germany

Karin Verspoor
Uni. Melbourne, Australia

Ondřej Bojar
Charles Uni., Czech Rep.

Arthur Boyer
LIMSI, CNRS,
Uni. Paris Saclay, France

Cristian Grozea
Fraunhofer Institute, Germany

Barry Haddow
Uni. Edinburgh, UK

Madeleine Kittner
Humboldt Uni., Germany

Yvonne Lichtblau
Humboldt Uni., Germany

Pavel Pecina
Charles Uni., Czech Rep.

Roland Roller
DFKI, Germany

Rudolf Rosa
Charles Uni., Czech Rep.

Amy Siu
MPI für Informatik, Germany

Philippe Thomas
DFKI, Germany

Saskia Trescher
Humboldt Uni., Germany

Abstract

Automatic translation of documents is an important task in many domains, including the biological and clinical domains. The second edition of the Biomedical Translation task in the Conference of Machine Translation focused on the automatic translation of biomedical-related documents between English and various European languages. This year, we addressed ten languages: Czech, German, English, French, Hungarian, Polish, Portuguese, Spanish, Romanian and Swedish. Test sets included both scientific publications (from the Scielo and EDP Sciences databases) and health-related news (from the Cochrane and UK National Health Service web sites). Seven teams participated in the task, submitting a total of 82 runs. Herein we describe the test sets, participating systems and results of both the automatic and manual evaluation of the translations.

1 Introduction

Automatic translation of texts allows readers to gain access to information present in documents written in a language in which the reader is not fluent. We identify two main use cases of machine translation (MT) in the biomedical domain: (a) making health information available to health professionals and the general public in their own language; and (b) assisting health professionals and researchers in writing reports of their research in English. In addition, it creates an opportunity for natural language processing (NLP) tools to be applied to domain-specific texts in languages for which few domain-relevant tools are available; i.e., the texts can be translated into a language for which there are more resources.

The second edition of the Biomedical Translation Task in the Conference for Machine Translation (WMT)[1] builds on the first edition (Bojar et al., 2016) by offering seven additional language pairs and new test sets. This year, we expanded to a total of ten languages in the biomedical task, namely, Czech (cs), German (de), English (en), French (fr), Hungarian (hu), Polish (pl), Portuguese (pt), Spanish (es), Romanian (ro) and Swedish (sv). Test sets included scientific publications from the Scielo and EDP Sciences databases and health-related news from Cochrane and the UK National Health Service (NHS).

Participants were challenged to build systems to enable translation from English to all other lan-

[1] http://www.statmt.org/wmt17/biomedical-translation-task.html

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 234–247
Copenhagen, Denmark, September 711, 2017. ©2017 Association for Computational Linguistics

guages, as well as from French, Spanish and Portuguese to English. We provided both training and development data but the teams were allowed to use additional in-domain or out-of-domain training data. After release of the test sets, the participants had 10 days to submit results (automatic translations) for any of the test sets and languages. We allowed up to three runs per team for each language pair and test sets.

We evaluated the submission both automatically and manually. In this work, we report details on the challenge, test sets, participating teams, the results they obtained and the quality of the automatic translations.

2 Training and test sets

We released test sets from four sources, namely, Scielo, EDP, Cochrane and NHS, as presented in Table 1. For training and development data, we referred participants to various biomedical corpora: (a) Biomedical Translation Corpora Repository[2], which includes titles from MEDLINE® and the Scielo corpus (Neves et al., 2016); (b) UFAL Medical Corpus,[3] which includes EMEA and PatTR Medical, among others; (c) development data from the Khresmoi project.[4] We provide details of the test sets below.

Scielo. Similar to last year, this dataset consisted of titles and abstracts from scientific publications retrieved from the Scielo database[5] and addressed the following language pairs: es/en, en/es, pt/en and en/pt. There were not enough articles indexed in 2017 with French titles or abstracts, so we relied on another source for en/fr and fr/en language pairs (namely, EDP as described below). Similar to last year, we crawled the Scielo site for publications containing both titles and abstracts in both English/Spanish or English/Portuguese language pairs. We considered only articles published in 2017 until that point (April/2017). We tokenized the documents using Apache OpenNLP[6] (with specific models for each language). The test set dataset was automatically created by aligning the GMA tool.[7] We manually checked the alignment of a sample and confirmed that around 88% of the sentences were correctly aligned.

EDP. Title and abstracts of scientific publications were collected from the open access publisher EDP Sciences[8] on March 15, 2017. The corpus comprises a selection of titles and abstracts of articles published in five journals in the fields of *Health* and *Life & Environmental Sciences*. The articles were originally written in French but the journals also publish the titles and abstracts in English, using a translation provided by the authors. The dataset was pre-processed for sentence segmentation using the Stanford CoreNLP toolkit[9] and aligned using YASA.[10] Manual evaluation conducted on a sample set suggests that 94% of the sentences are correctly aligned, with about 20% of the sentence pairs exhibiting additional content in one of the languages.

Cochrane and NHS. The test data was produced during the course of the KConnect[11] and HimL[12] projects. The test data contains health-related documents from Cochrane and NHS that were manually translated by experts from English to eight languages: cs, de, fr, hu, pl, ro, es and sv.

3 Participating teams and systems

We received submissions from seven teams, as summarized in Table 2. The teams came from a total of five countries (Germany, Japan, Poland, UK and USA) and from three continents. They include both research institutions and a company. An overview of the teams and their systems is provided below.

Hunter (Hunter College, City University of New York). The system from the Hunter College is based on Moses EMS, SRI-LM, GIZA++ (Xu et al., 2017). For the translation model, they generate word alignments using GIZA++ and mGIZA. For the language model, they relied on an interpolation of models that includes 6-grams with Kneser-Ney smoothing. Different corpora were used for the various languages

[2] https://github.com/biomedical-translation-corpora/wmt-task

[3] https://ufal.mff.cuni.cz/ufal_medical_corpus

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122

[5] http://www.scielo.org

[6] https://opennlp.apache.org/

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

[8] http://www.edpsciences.org

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

[10] http://rali.iro.umontreal.ca/rali/?q=en/yasa

[11] http://k-connect.org

[12] http://www.himl.eu/

Test sets	en/cs	en/de	fr/en	en/hu	pt/en	es/en	en/fr	en/pl	en/pt	en/es	en/ro	en/sv
Scielo					189/1897	158/1180			188/1806	158/1082		
EDP			85/699				84/750					
Cochrane	25/467	25/467	25/467	25/467		25/467	25/467	25/467		25/467	25/467	25/467
NHS	25/1044	25/1044	25/1044	25/1044		25/1044	25/1044	25/1044		25/1044	25/1044	25/1044

Table 1: Overview of the test sets. We present the number of documents and sentences in each test set.

Team ID	Institution
Hunter	Hunter College, City University of New York
kyoto	Kyoto University
Lilt	Lilt Inc.
LMU	Ludwig Maximilian University of Munich
PJIIT	Polish-Japanese Academy of Information Technology
uedin-nmt	University of Edinburgh
UHH	University of Hamburg

Table 2: Overview of the participating teams.

to which they submitted runs. The system was tuned using the WMT16 test sets (in the case of French and English) and on the HimL test sets for Cochrane and NHS. For training data, the team relied on a variety of corpora, depending on the language pair, which included MEDLINE, Europarl, Scielo, News Commentary, UFAL, EMEA, Cordis, among others.

kyoto (Kyoto University). The system from the team from Kyoto University is based on two previous papers (Cromieres et al., 2016; Cromieres, 2016). The participants describe it as a classic neural machine translation (NMT) system, however, we do not have further information regarding the datasets that have been used to train and tune the system for the WMT challenge.

Lilt (Lilt Inc.). The system from the Lilt Inc.[13] uses an in-house implementation of a sequence-to-sequence model with Bahdanau-style attention. The final submissions are ensembles between models fine-tuned on different parts of the available data.

LMU (Ludwig Maximilian University of Munich). LMU Munich has participated with an en2de NMT system (Huck and Fraser, 2017). A distinctive feature of their system is a linguistically informed, cascaded target word segmentation approach. Fine-tuning for the domain of health texts was done using in-domain sections of the UFAL Medical Corpus v.1.0 as a training corpus. The learning rate was set to 0.00001, initialized with a pre-trained model, and optimized using only the in-domain medical data. The HimL tuning sets were used for validation, and they tested separately on the Cochrane and NHS24 parts of the HimL devtest set.

PJIIT (Polish-Japanese Academy of Information Technology). PJIIT developed a translation model training, created adaptations of training settings for each language pair, and implemented byte pair encoding (BPE) (subword units) in their systems (Wolk and Marasek, 2017). Only the official parallel text corpora and monolingual models for the challenge evaluation campaign were used to train language models, and to develop, tune, and test their system. PJIIT explored the use of domain adaptation techniques, symmetrized word alignment models, the unsupervised transliteration models and the KenLM language modeling tool.

uedin-nmt (University of Edinburgh). The systems from the University of Edinburgh used a NMT trained with Nematus, an attentional encoder-decoder (Sennrich et al., 2017). Their setup follows the one from last year. This team again built BPE-based models with parallel and back-translated monolingual training data. New approaches this year included the use of deep architectures, layer normalization, and more compact models due to weight-tying and improvements in BPE segmentations.

UHH (University of Hamburg). All SMT models were developed using the Moses phrase-based MT toolkit and the Experiment Management System (Duma and Menzel, 2017). The preprocessing of the data consisted of tokenization, cleaning (6-80), lowercasing and normalizing punctuation. The tuning and the test sets were derived from WMT 2016 and WMT 2017. The SRILM toolkit

[13] https://lilt.com/

and Kneser-Ney discounting were used to estimate 5-gram language models (LM). For word alignment, GIZA++ with the default grow-diag-final-and alignment symmetrization method was used. Tuning of the SMT systems was performed with MERT. Commoncrawl and Wikipedia were used as general domain data for all language pairs except for EN/PT, where no Commoncrawl data was provided by WMT. As for the in-domain corpora, EMEA was used for all language pairs and Muchmore, ECDC, Pattr and Pubmed (all from UFAL Medical Corpus2) for the language pairs where data was available. The system made use of the training data provided by the previous Biomedical Task from 2016. The corpora corresponding to the general domain were concatenated into a single data source and the same procedure was applied for the in-domain corpora. This team investigated performing data selection for MT via Paragraph Vector and a Feed Forward Neural Network Classifier. Continuous distributed vector representations of the sentences were used as features for the classifier.

4 Evaluation

In this section, we present an overview of the submissions to the Biomedical Task and results in terms of both automatic and manual evaluation.

4.1 Submissions

An overview of the submissions is shown is Table 3. The participating teams submitted a total of 82 runs. No submissions were received for Swedish (en/sv) and Hungarian (en/hu).

4.2 Baselines

We provided baseline results only for the EDP and Scielo test sets, however, not for the other languages included in the Cochrane and NHS test sets.

Baseline. For the Scielo and EDP test sets, we compared the participants' results to our baseline system, which used the same approach as applied in last year's challenge (Bojar et al., 2016) for the evaluation of the Scielo dataset (Neves et al., 2016). The statistical machine translation (SMT) system used for the baseline was Moses (Koehn et al., 2007) with default settings. For es2en, en2es, fr2en, en2fr, pt2en and en2pt, the baseline system was trained as described in (Neves et al., 2016).

LIMSI baseline. For additional comparison, we also provided the results of an en2fr Moses-based system prepared by Ive et al. for their participation in the WMT16 biomedical track, which reflects the state of the art for this language pair (Ive et al., 2016a). The system uses in-domain parallel data provided for the biomedical task in 2016, as well as additional in-domain data[14] and out-of-domain data. However, we did not perform SOUL re-scoring.

4.3 Automatic evaluation

In this section, we provide the results for the automatic evaluation and rank the various systems based on those results. For the automatic evaluation, we computed BLEU scores at the sentence level using the multi-bleu and tokenization scripts as provided by Moses (`tokenizer` and `truecase`). For all test sets and language pairs, we compare the automatic translations to the reference one, as provided by each test set.

Results for the Scielo test sets are presented in Table 4. All three runs from the UHH team, for all four language pairs, obtained a much higher BLEU score than our baseline. However, this is not surprising given the simplicity of the methods used in the baseline system.

The BLEU scores for the EDP test set are presented in Table 5. While all system runs score above the baseline, only the Kyoto system outperforms the stronger baseline for en2fr. We rank the various submissions as follows:

- fr2en: Hunter (runs 1,2) < baseline < UHH (runs 1,2) < UHH (run 3) < kyoto (run 1).
- en2fr: baseline < Hunter (runs 1,2) < UHH (runs 1,2,3) < LIMSI baseline < kyoto (run 1) < kyoto (run 2).

The BLEU scores for the Cochrane test sets are presented in Table 6. The scores range from as low as 12.45 (for Polish) to as high as 48.99 (for Spanish). All scores were particularly high for Spanish (close to 50), but rather low for Polish and Czech (all below 30). While the BLEU value did not vary much for French (all around 30), these went from a range of 14 to 41 for Romanian. We rank the various submissions for each language as below:

[14] Cochrane translation corpus available at `http://www.translatecochrane.fr/corpus/` (Ive et al., 2016b)

Teams	en/cs	en/de	fr/en	pt/en	es/en	en/fr	en/pl	en/pt	en/es	en/ro
Hunter		CN	E2			C2NE2	CN			CN
kyoto			E			E2				
lilt		C2N2								
LMU		CN								
PJIIT	CN	CN					C3N3			CN
uedin-nmt	CN	CN					C2N2			C2N2
UHH		C3N3	E3	S3	S3	C3N3E3		S3	C3N3S3	

Table 3: Overview of submissions for each language pair and test set: [E]DP, [S]cielo, [C]ochrane and [N]HS. The number next to the letter indicates the number of runs that the team submitted for the corresponding test set.

Runs	pt/en	es/en	en/pt	en/es
baseline	36.35	31.50	30.52	27.31
UHH run1	43.84	37.14	39.14	36.08
UHH run2	43.93	37.47	39.38	35.93
UHH run3	43.88*	37.49*	39.21*	36.23*

Table 4: Results for the Scielo test sets. * indicates the primary run as identified by the participants.

Runs	fr/en	en/fr
baseline	17.47	12.32
LIMSI baseline	-	24.05
Hunter run1	15.10*	17.50*
Hunter run2	15.18	17.21
kyoto run1	25.21*	25.52
kyoto run2	-	27.04*
UHH run1	22.64	22.43
UHH run2	22.37	22.25
UHH run3	23.41*	22.79*

Table 5: Results for the EDP test sets. * indicates the primary run as declared by the participants.

- cs: PJIIT (run 1) < uedin-nmt (run 1).
- de: UHH (runs 1,2,3) < Hunter (run 1) < PIIJT (run 1) < lilt (run 1,2) < LMU < uedin-nmt (run 1).
- fr: Hunter (runs 1,2) < UHH (runs 1,2,3).
- pl: PIIJT (run 2) < Hunter (run 1) < PIIJT (runs 1,3) < uedin-nmt (run 2) < uedin-nmt (run 1).
- ro: Hunter (run 1) < PIIJT (run 1) < uedin-nmt (run 2) < uedin-nmt (run 1).

Finally, the BLEU scores for the NHS dataset are presented in Table 7. The scores range from as low as 10.56 (for Romanian, the lowest score across all test sets and languages) to as high as 41.22 (for Spanish). All scores were particularly high for Spanish (around 40), but rather low for Polish, Czech and Romanian (all below 30). We rank the various submissions for each language as shown below:

- cs: PJIIT (run 1) < uedin-nmt (run 1).
- de: UHH (runs 1,2,3) < Hunter (run 1) < PIIJT (run 1) < lilt (run 1,2) < LMU < uedin-nmt (run 1).
- fr: Hunter (run 1) < UHH (runs 1,2) < UHH (run 3).
- pl: PIIJT (run 2) < Hunter (run 1), PIIJT (runs 1,3) < uedin-nmt (run 2) < uedin-nmt (run 1).
- ro: Hunter (run 1) < PIIJT (run 1) < uedin-nmt (run 2) < uedin-nmt (run 1).

The BLEU values were generally lower for NHS than the ones obtained for the same teams for the Cochrane test sets. However, the rankings of systems and runs are nearly the same for the Cochrane and NHS test sets. The only exceptions were in French, where run 3 from UHH was higher than the others from the team, and for Polish, where the scores for Hunter and PIIJT (runs 1,3) were nearly the same.

4.4 Manual evaluation

We required teams to identify a primary run for each language pair, in the case that they submitted more than one run. These are the runs for which we performed manual evaluation. The following runs were considered to be primary: Hunter (run1), kyoto (run2 for en/fr, run1 for fr/en), lilt (run1), LMU (run1), PJIIT (run3 for pl, otherwise, run1), uedin-nmt (run1), UHH (run3).

We computed pairwise combinations of translations either between two automated systems, or one automated system and the reference translation. We compared all systems (primary) to the reference translation, as well as to each other system. We ran manual validation for all target languages and test sets. The human validators were

Cochrane	cs	de	fr	pl	es	ro
Hunter run1	-	24.72*	30.75*	17.16*	-	14.74*
Hunter run2	-	-	30.76	-	-	-
lilt run1	-	34.91*	-	-	-	-
lilt run2	-	33.97	-	-	-	-
LMU	-	36.44*	-	-	-	-
PJIIT run1	19.96*	25.13*	-	18.86	-	24.91*
PJIIT run2	-	-	-	12.45	-	-
PJIIT run3	-	-	-	18.88*	-	-
uedin-nmt run1	28.54*	37.11*	-	29.04*	-	41.18*
uedin-nmt run2	-	-	-	27.69	-	38.89
UHH run1	-	22.03	32.46	-	48.99	-
UHH run2	-	22.37	32.59	-	48.45	-
UHH run3	-	22.63*	33.16*	-	48.70*	-

Table 6: Results for the Cochrane test sets. * indicates the primary run as informed by the participants.

NHS	cs	de	fr	pl	es	ro
Hunter	-	20.45*	22.99*	14.09*	-	10.56*
lilt run1	-	27.57*	-	-	-	-
lilt run2	-	26.79	-	-	-	-
LMU	-	29.46*	-	-	-	-
PJIIT run1	15.93*	21.88*	-	14.32	-	18.10*
PJIIT run2	-	-	-	10.75	-	-
PJIIT run3	-	-	-	14.34*	-	-
uedin-nmt run1	22.79*	33.06*	-	23.15*	-	29.32*
uedin-nmt run2	-	-	-	19.87	-	27.32
UHH run1	-	18.71	31.79	-	40.97	-
UHH run2	-	19.80	31.89	-	41.20	-
UHH run3	-	19.66*	33.36*	-	41.22*	-

Table 7: Results for the NHS test sets. * indicates the primary run as informed by the participants.

native speakers of the languages and were either members of the participating teams or colleagues from the research community.

The validation task was carried out using the Appraise tool[15] (Federmann, 2010). For each pairwise comparison, we validated 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 translation had similar quality.
- A>B: when the quality of translation A was higher than B.
- Flag error: when the translations did not seem to be derived from the same input sentence. This is usually derived from error in the corpus alignment (for the Scielo and EDP datasets).

The manual validation for the Scielo test sets is presented in Table 8, for the comparison of the only participating team (UHH) to the reference translation. For en2es, the automatic translation scored lower than the reference one in 53 out of 100 pairs, but could still beat the reference translation in 23 pairs. For en2pt, the automatic translation was better only on 13 sentences pairs, while they could achieve similar quality to the reference translation on 31 cases. In the case of translations from Spanish or Portuguese to English, the reference scored better than the UHH around the same proportion, while the latter could only beat the reference in very few cases.

We present the results for the manual evaluation of the EDP 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:

- en2fr: Hunter < UHH < kyoto = reference
- fr2en: Hunter < UHH < kyoto < reference

Results for manual validation of the Cochrane test sets are presented in Table 10. We rank the various system as shown below:

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

Test set	Languages	Runs (A vs. B)	Total	A>B	A=B	A<B
Scielo	en2es	UHH vs. reference	100	23	24	53
	en2pt	UHH vs. reference	100	13	31	46
	es2en	UHH vs. reference	100	7	11	59
	pt2en	UHH vs. reference	100	10	20	50

Table 8: Results for the manual validation for the Scielo test sets. Values are absolute numbers (not percentages). They might not sum up to 100 due to the skipped sentences.

Test set	Languages	Runs (A vs. B)	Total	A>B	A=B	A<B
EDP	en2fr	UHH vs. reference	100	3	4	87
		UHH vs. Hunter	100	42	46	7
		UHH vs. kyoto	100	10	21	64
		Hunter vs. reference	100	0	2	93
		kyoto vs. reference	100	28	30	35
		Hunter vs. kyoto	100	3	10	82
	fr2en	UHH vs. reference	100	5	9	72
		UHH vs. Hunter	100	79	5	10
		UHH vs. kyoto	100	26	7	62
		Hunter vs. reference	100	2	4	79
		kyoto vs. reference	100	25	9	48
		Hunter vs. kyoto	100	3	9	81

Table 9: 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.

- cs: PIIJT < uedin-nmt < reference
- de: UHH < Hunter = PJIIT < Lilt < LMU < uedin-nmt = reference
- fr: UHH < Hunter < reference
- pl: Hunter = PIIJT < uedin < reference
- es: UHH < reference
- ro: Hunter < PIIJT < uedin < reference

Results for manual validation of the NHS test sets are presented in Table 11. We rank the various system as shown below:

- cs: PIIJT < uedin-nmt < reference
- de: Hunter = UHH < PIIJT < Lilt < LMU = uedin-nmt < reference
- fr: UHH < Hunter < reference
- pl: Hunter < PIIJT < uedin < reference
- es: UHH < reference
- ro: Hunter < PIIJT < uedin < reference

For the Polish language in the NHS test set, the evaluator skipped too many sentences (68 out of 100) to enable a comparison between Hunter and PIIJT. However, we ranked the PIIJT system higher than Hunter given that the former scored 21 times better that the latter (in contrast to 7). However, there is inadequate data to support assigning a clear difference between the two systems. Indeed, both systems have similar quality for this language in the Cochrane test set.

5 Discussion

In this section we present, for each target language, some insights from the automatic validation, the quality of the translations, as well as future work that we plan to implement in the next edition of the challenges.

5.1 Performance of the systems

The results obtained by the teams show interesting point of discussion regarding the impact of methods and amount of training data. Considering all the results in Tables 4-7, the highest BLEU score (48.99) of all runs across all test sets was obtained by the UHH system for en2es (Cochrane test set). The same team also scored high (above 40) for the NHS en2es test set and for the Scielo pt2en test set. The only other team that obtained BLEU scores in the same range (above 40) was uedin-nmt for the Cochrane en2ro test set.

No automatic system was able to outperform or match the reference translations on manual evaluation; hence the automated systems all still have room for improvement. Interestingly, it can be noted that the best performing system on the EDP

Test set	Languages	Runs (A vs. B)	Total	A>B	A=B	A<B
Cochrane	cs	PIIJT vs. reference	100	4	1	95
		PIIJT vs. uedin-nmt	100	19	6	75
		uedin-nmt vs. reference	100	8	38	54
	de	Hunter vs. reference	100	5	12	83
		Hunter vs. Lilt	100	12	20	68
		Hunter vs. LMU	100	6	20	73
		Hunter vs. PJIIT	100	26	41	33
		Hunter vs. uedin-nmt	100	3	12	85
		Hunter vs. UHH	100	42	30	28
		Lilt vs. reference	100	19	22	59
		LMU vs. reference	100	17	32	51
		PJIIT vs. reference	100	2	8	90
		uedin-nmt vs. reference	100	31	29	40
		UHH vs. reference	100	93	6	1
		Lilt vs. LMU	100	23	24	50
		Lilt vs. PJIIT	100	66	19	15
		Lilt vs. uedin-nmt	100	14	22	63
		Lilt vs. UHH	100	81	8	11
		LMU vs. PJIIT	100	82	9	7
		LMU vs. uedin-nmt	100	19	50	31
		LMU vs. UHH	100	82	10	3
		PJIIT vs. uedin-nmt	100	14	22	64
		PJIIT vs. UHH	100	34	44	22
		uedin-nmt vs. UHH	100	87	5	8
	fr	UHH vs. reference	100	8	8	83
		UHH vs. Hunter	100	8	51	40
		Hunter vs. reference	100	11	10	79
	pl	Hunter vs. PJIIT	100	43	7	48
		Hunter vs. reference	100	4	8	88
		Hunter vs. uedin-nmt	100	16	0	84
		PJIIT vs. reference	100	3	11	86
		PJIIT vs. uedin-nmt	100	16	4	80
		uedin-nmt vs. reference	100	15	34	51
	es	UHH vs. reference	100	4	29	67
	ro	Hunter vs. PJIIT	100	6	20	74
		Hunter vs. reference	100	1	3	96
		Hunter vs. uedin-nmt	100	5	8	87
		PJIIT vs. reference	100	3	6	91
		PJIIT vs. uedin-nmt	100	20	21	59
		uedin-nmt vs. reference	100	4	32	64

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

Test set	Languages	Runs (A vs. B)	Total	A>B	A=B	A<B
NHS	cs	PIIJT vs. reference	100	4	20	76
		PIIJT vs. uedin-nmt	100	28	23	49
		uedin-nmt vs. reference	100	7	41	52
	de	Hunter vs. reference	100	0	9	91
		Hunter vs. Lilt	100	28	29	43
		Hunter vs. LMU	100	17	12	68
		Hunter vs. PJIIT	100	32	28	40
		Hunter vs. uedin-nmt	100	12	18	70
		Hunter vs. UHH	100	34	36	30
		Lilt vs. reference	100	2	35	63
		LMU vs. reference	100	4	30	62
		PJIIT vs. reference	100	1	24	74
		uedin-nmt vs. reference	100	5	45	46
		UHH vs. reference	100	2	18	79
		Lilt vs. LMU	100	19	44	33
		Lilt vs. PJIIT	100	46	24	30
		Lilt vs. uedin-nmt	100	11	23	66
		Lilt vs. UHH	100	47	28	25
		LMU vs. PJIIT	100	56	22	18
		LMU vs. uedin-nmt	100	37	27	33
		LMU vs. UHH	100	59	19	18
		PJIIT vs. uedin-nmt	100	8	24	68
		PJIIT vs. UHH	100	51	21	28
		uedin vs. UHH	100	63	29	8
	fr	UHH vs. reference	100	0	2	98
		UHH vs. Hunter	100	6	27	67
		Hunter vs. reference	100	11	23	65
	pl	Hunter vs. PJIIT	100	7	4	21
		Hunter vs. reference	100	14	2	84
		Hunter vs. uedin-nmt	100	8	11	48
		PJIIT vs. reference	100	9	8	83
		PJIIT vs. uedin-nmt	100	8	16	62
		uedin-nmt vs. reference	100	11	14	75
	es	UHH vs. reference	100	1	32	67
	ro	Hunter vs. PJIIT	100	10	38	52
		Hunter vs. reference	100	1	7	92
		Hunter vs. uedin-nmt	100	4	27	62
		PJIIT vs. reference	100	3	16	81
		PJIIT vs. uedin-nmt	100	24	34	41
		uedin-nmt vs. reference	100	6	26	68

Table 11: Results for the manual validation for the NHS test sets. Values are absolute numbers (not percentages). They might not sum up to 100 due to the skipped sentences.

en2fr dataset (Kyoto) compared very favorably to the reference and was found to be equal to or better than the reference in 62% (58/93) of the manually evaluated sentences. In general, the kyoto and uedin-nmt systems seemed to consistently outperform other competitors.

Regarding comparison of results to the ones obtained in the last year's edition of the challenge, we can only draw conclusions for the Scielo test set. The only participating team (UHH) obtained much higher BLEU scores for en2pt (39 vs. 19), pt2en (43 vs. 21) and es2en (37 vs. 30). However, results for en2es were just a little higher than last year's ones (36 vs. 33).

As the performance of the methods improves on the biomedical domain, it will make sense to introduce additional domain-oriented evaluation measures that provide a document-level assessment focused on the clinical validity of the translations, rather than the grammatical correctness and fluency.

5.2 Best-performing methods

For languages which received submissions from several systems, such as en2de over Cochrane and NHS data, the systems based on neural networks (e.g., uedin-nmt and LMU) performed substantially better than those based on SMT (e.g., UHH and Hunter). In many runs, the difference in BLEU score was greater than 10 points. The superiority of NMT systems was also observed in the EDP test set, as implemented in the Kyoto system. However, we also note that a state-of-the-art statistical system relying on rich in-domain and out-of-domain data still performs well (as seen in the strong results of the LIMSI system).

Finally, some teams submitted more than one run but we only observed significant differences in BLEU scores in a few cases, namely, kyoto (EDP en2fr test set), PJIIT (Cochrane/NHS pl test set), uedin (Cochrane/NHS pl and ro test sets). In the case of the PJIIT systems, the best performing one is an extended version of the base SMT system that includes domain adaptation, among other additional features. In the case of the uedin-nmt system, the best performing run relied on advanced techniques, such as +right-to-left re-ranking.

5.3 Differences across languages

Even if some teams relied on equal or similar methods for the different languages, the same system might perform better for certain languages then for others. This is probably due to amount (or quality) of training data available for each language and also due to different linguistic properties of the language pair in question.

For instance, the UHH team developed a SMT system which was trained on a variety of domain and out-of-the-domain data. This system achieved good performance for English, Portuguese and Spanish (around 30-48), but their results for German were much poorer (around 18-22). Indeed, the system obtained the lowest rank position for German for the Cochrane and NHS test sets. The participants report that this is probably due to the amount of training data available for this language (personal communication), even though other teams could obtain much higher BLEU scores for those same test sets, e.g., up to 37 points in the case of the uedin-nmt system.

Such differences across languages was also observed for other systems (higher than 10-20 points in the BLEU score). For instance, scores for the uedin-nmt system ranged from 22 (for Czech) to 41 (for Romanian). Interestingly, the scores for the Hunter system ranged from 10 (for Romanian, in contrast to higher scores from uedin-nmt system) to 30 (for French). The Hunter team seems to have used the same approach across all languages and all of these were trained on a variety of corpora. On the other hand, the uedin-nmt team seems to have used slightly different network architectures for each language (Sennrich et al., 2017).

5.4 Differences across datasets

Given that the methods and corpora seem to be largely the same for a particular language, differences in BLEU scores across the test sets are probably related to the the characteristics of these. Few teams submitted runs for more than one test set and only one team (UHH) submitted runs for all test set (for one particular language).

For Spanish, the UHH team obtained considerable differences in BLEU score for Scielo (around 36), NHS (around 41) and Cochrane (around 48). However, their system paper does not give much insight on the reason for such differences (Duma and Menzel, 2017). We can hypothesize that lower scores in the Scielo datasets are due to the fact that the reference translation is not a perfect translation of the source document and sentence alignment was performed automatically.

For French, the Hunter team obtained lower scores in the EDP test set (around 17) and higher ones in the NHS (almost 23) and Cochrane test sets (around 30). Similarly, the UHH team obtained lower scores for the EDP (around 22) and higher ones for Cochrane and NHS (around 31-32). The reason for these differences is probably the same as for the Scielo test set: this is an automatically acquired test set, whose documents were automatically aligned. While the quality of the automatic alignment is high (estimated at 88% accuracy for Scielo and 94% for EDP), we can also note that the translations in these test sets are created by the authors of the articles who are neither professional translators nor native speakers of the all the languages involved.

On the other hand, differences also occurred between the Cochrane and NHS test sets, although these were manually translated by professionals. Such differences were small for most systems (24 vs. 20 for Hunter, 22 vs. 19 for UHH, 25 vs. 21 for PIIJT), for German in the Cochrane and NHS test sets, respectively. However, some cases show larger differences, such as the uedin-nmt system for Romanian (41 vs. 29 for Cochrane and NHS, respectively). We observed that that the average sentence length is higher for Cochrane (with some very long sentences included) while there are many short sentence fragments in the NHS test set. However, both can be problematic for MT as this can scramble long sentences, and trip up over sentence fragments since most of the training data consists of full sentences.

5.5 Differences between manual and automatic evaluations

We checked for differences between the manual and automatic evaluations, i.e., whether a team performed better than another in the manual evaluation but the other way round in the automatic evaluation. We observed small differences for Polish (Cochrane and NHS test sets) between the Hunter and PIIJT teams, but these are probably not significant and both systems have probably similar performance. We observed the same for the UHH and Hunter systems for German (NHS test set). However, we found a more interesting contradiction between Hunter and UHH systems for French in both Cohrane and NHS test sets. UHH obtained higher BLEU scores than Hunter (32-33 vs. 30 and 31-33 vs. 23, for Cochrane and NHS, respectively). However, in the manual evaluation, our expert chose Hunter as being better than UHH in many more sentences (40 vs. 8 and 67 vs. 6, respectively).

5.6 Quality of the automatic translations

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

Czech: The outputs of the weaker system, PIIJT, were rather unsurprising, featuring a wide range of well-known issues of phrase-based SMT, including inflection errors that violate both long-distance and short-distance morphological agreement, errors in missing or surplus negation, untranslated and uninflected rare words, wrong disambiguation of word meanings, etc. On the other hand, the quality of the neural uedint-nmt system is remarkably better, with no negation errors spotted, agreement errors generally limited to long-distance dependencies, only rare disambiguation errors (often domain-specific, e.g. "drug", "study", "review"), and a much bolder attempt at handling unknown or rare words. On one hand, we spotted cases where it would have been better to leave the word untranslated, or to only perform modest transliterations, as in "haemoglobin", which is similar enough to the "hemoglobin" used in Czech to be understandable as it is, but got translated to "hemoroidy" ("hemorrhoids") instead; on the other hand, both correct and incorrect translations of rare words were nearly always correctly inflected. Occasionally, we also noticed a missing or surplus word – especially with auxiliaries, such as reflexive pronouns or forms of the verb "be".

English: Overall, the assessor found the quality of translations into English improved from 2016. Some of the problems observed in the prior year persisted, including inappropriate capitalization of terms (terms were capitalized although they were neither proper nouns nor acronyms) for some translations. Other issues such as incorrect word order as well as untranslated and missing words were observed. Especially in fr2en translations, incorrect word order occurred when the noun-before-adjective grammar in French was erroneously preserved in English; for instance, "douleur oro-faciale" was translated as "pain oro-facial". Sometimes, however, untranslated words could still be deciphered because the French words were similar to the English equivalents, such as "biomatériaux" vs. "biomaterials", and "tolérance immunologique" vs. "immunological tolerance". As for missing words, translations were severely impacted when entire phrases were omitted, for instance when two consequences of a procedure were reduced to only one.

French: The quality of translations varied from poor to good. The issues that we encountered were similar to last year and included grammatical errors such as incorrect subject/verb or adjective/noun agreement, untranslated passages, incorrect lexical choice due to a lack of word sense disambiguation. One recurring mistake was the translation of the term "female" as "femelle", which is appropriate for animals instead of "femme", which is appropriate for humans. This year, the best systems showed an ability to successfully translate some acronyms. However, complex hyphenated terms remained challenging (for example, "38-year-old", "mid-60s", "immunoglobulin-like").

German: Overall, the quality of translations to German ranges from very good to poor. Comparing between the different systems, the translation with the better syntax, grammar and use of technical terms was preferred. When both translations were equally bad their performance was assigned equal. Poor translations are mostly characterized by incorrect syntax and grammar. Syntactic errors are usually due to missing predicates, the usage of two or more predicates in one sentence, and strange word order, especially in long sentences. This often led to confusion or even not understanding the meaning of a sentence. Usual grammar errors included incorrect conjugation of verbs as "wir suchte" instead of "wir suchten" (we searched). In well performing systems syntax and grammar are often correct. Their difference to the reference is often due to not using the most appropriate word. This does not influence the meaning of the sentence. Only as a native speaker one would rather use a different word. All systems seem to have problems with certain technical terms. Usually this occurs when the German translation is very different from the English term. For instance, "to restart a person's heart" is often word-by-word translated into "Neustart des Herzen" while in German this procedure is called "Reanimation des Herzens". The pairwise evaluation of the two best performing teams (LMU and uedin-nmt) indicates, that they often provide similar sentences in terms of grammar and token order.

Portuguese: Only one team (UHH) submitted translation for Portuguese (Scielo dataset). In comparison to submissions from the previous challenge (Bojar et al., 2016), we found the quality of the translations considerably better. As expected, longer sentences usually contained more mistakes and were harder to understand than shorter sentences, usually due to the wrong placement of the commas and conjunctions (e.g., and). For instance, the translation "diâmetro tubular, altura do epitélio seminífero e integridade" was derived from the English version of the reference clause "diâmetro dos túbulos seminíferos, altura e integridade do epitélio seminífero". However, the same can be also stated for some reference sentences, which could have a higher quality.

Regarding more common mistakes, we observed missing articles, such as "Extratos vegetais" versus "Os extratos das espécies vegetais". However, we observed fewer instances of untranslated English words in comparison to last year, which seems to indicate a better coverage of the biomedical terminology. In some sentences, such cases were observed for terms which were skipped by the translation system, such as "método de manometria de alta resolução" for "high-resolution manometry method for esophageal manometry". The same mistake was observed for acronyms, e.g., DPS (death of pastures syndrome) instead of SMP (síndrome da morte das pastagens). However, we also found correct translations for acronyms, e.g., SII (síndrome do intenstino irritavel) instead of IBS (irritable bowel syndrome). Finally, we observed other minor mistakes: (a) nominal concordance, e.g., "O fortalecimento muscular progressiva"; (b) wrong word ordering, e.g., "plantadas áreas florestais" instead of "áreas florestais plantadas"; (c) wrong verb tense, e.g., "coeficiente de correlação linear de Pearson spearmans determinado" instead of "determinou"; (d) wrong verb conjugation, e.g, "a umidade relativa, temperatura, velocidade do vento e intensidade de luz foi...", instead of "foram"; and (e) no contraction when necessary, e.g., "em as" instead of "nas".

Spanish: Compared to last year's challenge translations, the quality of the translations into Spanish is significantly better. Despite some small variations, many of the produced translations are valid translations of the original text. There are still cases in which there are mistakes such as with verb tenses “a menudo oír voces”, which should be “a menudo oyen voces”. There are translations with similar meaning but not entirely the same meaning such as “hace aparecer” vs. the reference translation “ocurren”. In some cases, there are some incorrect phrases such as “teléfono NHS informar sobre” vs. the reference translation “llame por teléfono el sistema informativo de NHS en”. Translation systems seem to have better alignment between masculine/feminine and singular/plural articles as compared to last year. In addition, the number of missing words is lower in the Spanish submissions.

Romanian: The quality varied from good translations to clearly underperforming ones. When both translations were good, the one that was grammatically correct was preferred. When one used an awkward language or did not use domain-specific terms such as “traumatism cranian” or “presiune intracraniana”, the other one was preferred. We noticed that these translations can be very dangerous, especially when the form is good (and thus the appearance of quality is high). For instance, in one case, “vasopressor” was translated as “vasodilatatoare”, which is the precise antonym. A frequent mistake was the translation of “trials” as “procese”, which would have been correct for “law suits” but not for clinical trials. Somewhat confusing was translating “norepinephrine” as “noradrenaline”, as they look different but are two names of the same substance. For the bad and very bad translations, errors abounded up to the point that both were equally useless and therefore marked as equal (in the sense of equally bad); this happened quite often. In general, we preferred translations that did not mislead and were still possible to understand despite their many flaws. Among the frequent translation errors, we identified the following: untranslated words, grammatical errors (case, gender), random characters and even Cyrillic (for no apparent reason) and context which were frequently not considered (e.g. “shots” translated to “gloante” and “impuscaturi”, those words having to do with weapons not with syringes). Other strange errors included unrelated words from other fields, especially “subcontractantul copolimerului” or “transductoare AFC”.

6 Conclusions

We presented the results of the second edition of the Biomedical task in the Conference for Machine Translation. The shared task addressed a total of ten languages and received submission from seven teams. In comparison to last year, we observed an increase on the performance of the systems in terms of higher BLEU scores as well as an improvement in the quality of the translations, as observed during manual validation. The methods used by the systems included statistical and neural machine translation techniques, but also incorporated many advanced features to boost the performance, such as domain adaptation.

Despite the comprehensive evaluation that we show here, there is still room for improvement in our methodology. All by professionals were rather small (up to 1000 sentences), which means that some of the our conclusions might not hold on a larger benchmark. Further, we did not perform statistical tests when ranking the various systems and runs in both manual and automatic evaluations. Furthermore, each combination of two translations or one translation and reference was evaluated by a single expert, given the high number of submissions and the difficulty of finding available experts. On the other hand, most results obtained through manual validation were consistent with the automatic validation, suggesting that automatic scoring is sufficiently meaningful.

Acknowledgments

We would like to thank the support of participants and colleagues in the manual validation of the translations. We thank Julia Ive and Franck Burlot for their assistance with the LIMSI translation system. The Cochrane and NHS test sets were provided by the EU H2020 projects HimL (contract no. 644402) and KConnect (contract no. 644753).

References

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.

Fabien Cromieres. 2016. Kyoto-NMT: a Neural Machine Translation implementation in Chainer. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 307–311.

Fabien Cromieres, Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2016. Kyoto University Participation to WAT 2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 166–174.

Mirela-Stefania Duma and Wolfgang Menzel. 2017. Automatic Threshold Detection for Data Selection in Machine Translation. In *Proceedings of the Second Conference on Machine Translation (WMT17) at the Conference on Empirical Methods in Natural Language Processing*.

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

Matthias Huck and Alexander Fraser. 2017. Lmu Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proceedings of the Second Conference on Machine Translation (WMT17) at the Conference on Empirical Methods in Natural Language Processing*.

Julia Ive, Aurélien Max, and François Yvon. 2016a. Limsi's contribution to the wmt'16 biomedical translation task. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 469–476. http://www.aclweb.org/anthology/W/W16/W16-2337.

Julia Ive, Aurélien Max, François Yvon, and Philippe Ravaud. 2016b. Diagnosing high-quality statistical machine translation using traces of post-edition operations. In *Proceedings of the LREC 2016 Workshop: Translation evaluation From fragmented tools and data sets to an integrated ecosystem*. European Language Resources Association (ELRA), Portorož, Slovenia, pages 55–62. http://www.cracking-the-language-barrier.eu/wp-content/uploads/LREC-2016-MT-Eval-Workshop-Proceedings.pdf.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180. http://dl.acm.org/citation.cfm?id=1557769.1557821.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation (WMT17) at the Conference on Empirical Methods in Natural Language Processing*.

Krzysztof Wolk and Krzysztof Marasek. 2017. PJIIT's systems for WMT 2017 Conference. In *Proceedings of the Second Conference on Machine Translation (WMT17) at the Conference on Empirical Methods in Natural Language Processing*.

Jia Xu, Yi Zong Kuang, Shondell Baijoo, Jacob Lee, Mir Ahmed, Uman Shahzad, Meredith Lancaster, and Chris Carlan. 2017. Supervised Study for Young Machine Translators: Hunter Machine Translation Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation (WMT17) at the Conference on Empirical Methods in Natural Language Processing*.