WMT 2021 - Biomedical task

August 5, 2021

1 Results for Automatic Evaluation

In all result tables, an asterisk * indicates the primary run, as informed by the participants, in the case of multiple runs.

1.1 Terminology test set

For the evaluation of terminology we provide two metrics for the en_eu task: (i) accuracy, by relying on strict matches (case insensitive) between ground truth and predictions; and (ii) BLEU score, as measured by the NLTK module sentence_bleu.

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

Table 1: Performance scores for the Terminology test set (en2eu).

1.2 Abstract test set

BLEU scores were calculated using the multi-eval tool and tokenization as provided in Moses.

Teams	**Runs**	**BLEU**
FJDMATH	run1	0.1453
	run2*	0.1403

Table 2: Performance scores for the Abstract test set (en2eu).

1.3 MEDLINE test sets

BLEU scores were calculated using the multi-eval tool and tokenization as provided in Moses.

Teams	**Runs**	**en2de**	**en2es**	**en2fr**	**en2it**	**en2pt**	**en2ru**	**en2zh**
Huawei_AGI	run1	0.2657	-	0.4465	0.3638	-	-	0.4278
	run2	0.2674	-	0.4351	0.3666	-	-	0.4367
	run3*	0.2657	-	0.4427	0.3750	-	-	0.4237
Huawei_TSC	run1	0.2711	-	-	-	-	-	0.4567
	run2	0.2776	-	-	-	-	-	0.4569
	run3	0.2711	-	-	-	-	-	0.4578
LISN	run1*	-	-	0.3839	-	-	-	-
	run2	-	-	0.3843	-	-	-	-
	run3	-	-	0.4250	-	-	-	-
NeMo	run1	-	-	-	-	-	0.3011	-
	run2*	-	-	-	-	-	0.3078	-
talp_upc	run1	-	0.3899	-	-	-	-	-
	run2*	-	0.3899	-	-	-	-	-
TMT	run1	0.2332	0.4157	0.4390	-	-	0.2543	-
Transperfect	run1	-	0.4852	-	-	-	0.2798	0.3972
	run2*	-	0.4742	-	-	-	0.2664	0.3966
	run3	-	0.4666	-	-	-	-	-
Volctrans	run1	-	-	-	-	-	-	0.4335
	run2	-	-	-	-	-	-	0.4357
	run3*	-	-	-	-	-	-	0.4291
ZengHuiMT	run1	-	-	-	-	-	-	0.4065

Table 3: BLEU scores for all test sentences, from English. For the Volctrans team, we renamed the runs: run1=run1, run2=nnmt, run3=nnmtne.

Teams	**Runs**	**de2en**	**es2en**	**fr2en**	**it2en**	**pt2en**	**ru2en**	**zh2en**
Huawei_AGI	run1	0.3430	-	0.4745	0.4176	-	-	0.3931
	run2	0.3507	-	0.4754	0.4172	-	-	0.3788
	run3*	0.3484	-	0.4755	0.4156	-	-	0.3937
Huawei_TSC	run1	0.3635	-	-	-	-	-	0.3825
	run2	0.3650	-	-	-	-	-	0.3870
	run3	0.3669	-	-	-	-	-	0.3862
LISN	run1	-	-	0.4248	-	-	-	-
	run2	-	-	0.4039	-	-	-	-
	run3*	-	-	0.4208	-	-	-	-
MT Learner	run1*	-	-	-	0.4157	0.5515	0.3548	-
	run2	-	-	-	0.4156	0.5617	0.3470	-
NeMo	run1	-	-	-	-	-	0.3623	-
nrpu-fjwu	run1*	0.3103	0.4489	0.3768	-	-	-	-
	run2	0.3087	0.4508	0.3825	-	-	-	-
	run3	0.2975	0.4518	0.3679	-	-	-	-
talp_upc	run1	-	0.4112	-	-	-	-	-
	run2*	-	0.4112	-	-	-	-	-
TMT	run1	0.3816	0.5299	0.4805	-	-	0.2998	-
Transperfect	run1	-	0.5152	-	-	-	0.3496	0.3275
	run2*	-	0.4909	-	-	-	0.3415	0.3200
	run3	-	0.4884	-	-	-	-	-
Volctrans	run1	-	-	-	-	-	-	0.2829
	run2	-	-	-	-	-	-	0.3711
ZengHuiMT	run1	-	-	-	-	-	-	0.2826

Table 4: BLEU scores for all test sentences, into English. For the Volctrans team, we renamed the runs: run1=base, run2=nnmt.

Teams	**Runs**	**en2de**	**en2es**	**en2fr**	**en2it**	**en2pt**	**en2ru**	**en2zh**
Huawei_AGI	run1	0.3172	-	0.4531	0.4301	-	-	0.4342
	run2	0.3198	-	0.4424	0.4334	-	-	0.4440
	run3*	0.3172	-	0.4489	0.4425	-	-	0.4293
Huawei_TSC	run1	0.3259	-	-	-	-	-	0.4639
	run2	0.3329	-	-	-	-	-	0.4640
	run3	0.3259	-	-	-	-	-	0.4650
LISN	run1*	-	-	0.3912	-	-	-	-
	run2	-	-	0.3913	-	-	-	-
	run3	-	-	0.4293	-	-	-	-
NeMo	run1	-	-	-	-	-	0.4139	-
	run2*	-	-	-	-	-	0.4112	-
talp_upc	run1	-	0.4084	-	-	-	-	-
	run2*	-	0.4142	-	-	-	-	-
TMT	run1	0.2765	0.4354	0.4456	-	-	0.3289	-
Transperfect	run1	-	0.5117	-	-	-	0.3686	0.4029
	run2*	-	0.5012	-	-	-	0.3492	0.4025
	run3	-	0.4917	-	-	-	-	-
Volctrans	run1	-	-	-	-	-	-	0.4406
	run2	-	-	-	-	-	-	0.4433
	run3*	-	-	-	-	-	-	0.4361
ZengHuiMT	run1	-	-	-	-	-	-	0.4126

Table 5: BLEU scores for ”OK” aligned test sentences, from English. For the Volctrans team, we renamed the runs: run1=run1, run2=nnmt, run3=nnmtne.

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

Table 6: BLEU scores for ”OK” aligned test sentences, into English. For the Volctrans team, we renamed the runs: run1=base, run2=nnmt.