WMT 2022 - Biomedical task

October 13, 2022

1 Results for the sub-task ClinSpEn

The ClinSpEn sub-task consists of three different tracks: (i) clinical cases (Table 1); (ii) clinical terminology (Table 2); and (iii) ontology concepts (Table 3). Each track was evaluated on five metrics: COMET, METEOR, SacreBLEU, BLEU and ROUGE, with the main one being SacreBLEU. Participants were allowed up to 7 different runs.

1.1 ClinSpEn Clinical Cases

Teams	**Runs**	**COMET**	**METEOR**	**ROUGE**	**BLEU**	**SacreBLEU**
Avellana Translation	run1	0.3920	0.6437	0.6333	0.3519	36.64
DtranX	run1	0.4610	0.6633	0.6490	0.3926	41.06
Logrus_UoM	run1	0.4237	0.6337	0.6270	0.3650	38.17
Optum	run1	0.4208	0.6310	0.6134	0.3508	36.89
	run2	0.4233	0.6353	0.6214	0.3577	37.47
	run3	0.4349	0.6405	0.6241	0.3594	37.65
	run4	0.4425	0.6447	0.6285	0.3642	38.12

Table 1: Results for the Clinical Cases track

1.2 ClinSpEn Clinical Terminology

Teams	Runs	COMET	METEOR	ROUGE	BLEU	SacreBLEU
Avellana Translation	run1	0.1966	0.5706	0.6864	0.1565	15.88
DtranX	run1	1.1152	0.6109	0.7013	0.3521	35.84
Huawei	run1	1.0617	0.6004	0.6884	0.3044	30.81
	run2	1.0618	0.6004	0.6884	0.3044	30.81
	run3*	0	0	0	0	0
	run4	0.7491	0.5200	0.5964	0.2054	20.87
	run5	0.8648	0.5444	0.6257	0.2560	26.09
	run6	1.1447	0.6096	0.7057	0.3635	36.76
	run7	1.1902	0.6240	0.7218	0.4132	41.57
Logrus_UoM	run1	0.9791	0.5884	0.6719	0.2667	26.87
Optum	run1*	0	0	0	0	0
	run2	0.9824	0.5742	0.6567	0.2757	27.9445

Table 2: Results for the Clinical Terminology track. Runs marked with an asterisk (*) had a submission format problem.

1.3 ClinSpEn Ontology Concepts

Teams	Runs	COMET	METEOR	ROUGE	BLEU	SacreBLEU
Avellana Translation	run1	0.3841	0.5707	0.7621	0.3042	31.72
DtranX	run1	1.2496	0.6275	0.7839	0.5724	58.24
Logrus_UoM	run1	0.9494	0.6261	0.7688	0.3674	39.10
Optum	run1	1.1197	0.5880	0.7479	0.4396	44.97
	run2	0	0	0	0	0
	run3	1.0914	0.6072	0.7511	0.4294	43.83
	run4	1.1062	0.5767	0.7339	0.4194	42.59

Table 3: Results for the Ontology Concepts track. Runs marked with an asterisk (*) had a submission format problem.

2 Results for Automatic Evaluation of the MEDLINE test sets

We separate results into three sections: (i) results for submissions to the WMT submission system; (ii) results for submissions to the biomedical test sets in OCELoT; and (iii) results for the submissions to the general Task in OCELoT, which includes some of the biomedical test sets.

BLEU scores were calculated using the multi-eval tool and tokenization as provided in Moses. In all result tables, an asterisk * indicates the primary run, as informed by the participants, in the case of multiple runs.

2.1 WMT Submission system

Teams	**Runs**	**en2de**	**en2es**	**en2fr**	**en2it**	**en2pt**	**en2ru**	**en2zh**
ChicHealth	run1	-	-	-	-	-	-	0.5571
ECNU-MT	run1	-	-	-	-	-	-	0.3985
HuaweiTSC	run1	0.39	-	0.4017	-	-	0.4127	0.5079
	run2	0.3914	-	0.3881	-	-	0.4053	0.5078
	run3	0.3891	-	0.39	-	-	0.4063	0.5068
Huawei-BabelTar	run1	0.3342	0.447	0.3785	0.4649	0.5255	0.3697	0.4768
	run2	0.3313	0.4415	0.3749	0.4783	0.5174	0.3674	0.473
	run3	0.3304	0.4475	0.3621	0.4848	0.5147	0.3703	0.4513
PAHT	run1	-	-	-	-	-	-	0.4826
SRT	run1	-	0.5214	-	-	-	-	-
	run2	-	0.5196	-	-	-	-	-
	run3	-	0.5235	-	-	-	-	-
Baseline	-	0.2943	0.3915	0.2812	0.4713	0.4239	0.2759	0.3979

Table 4: BLEU scores for "OK" aligned test sentences, from English.

Teams	**Runs**	**de2en**	**es2en**	**fr2en**	**it2en**	**pt2en**	**ru2en**	**zh2en**
ChicHealth	run1	-	-	-	-	-	-	0.3427
	run2	-	-	-	-	-	-	0.3648
	run3	-	-	-	-	-	-	0.4614
ECNU-MT	run1	-	-	-	-	-	-	0,2475
	run2	-	-	-	-	-	-	0.2449
HuaweiTSC	run1	0.4695	-	0.5095	-	-	0.4886	0.4269
	run2	0.4712	-	0.5036	-	-	0.5001	0.4256
	run3	0.4682	-	0.5048	-	-	0.4958	0.4276
Huawei-BabelTar	run1	0.431	0.566	0.4908	0.4883	0.5603	0.4616	0.4612
	run2	0.4375	0.5902	0.4886	0.4916	0.5544	0.4626	0.4249
	run3	0.4338	0.5864	0.4936	0.4989	0.5563	0.4675	0.418
PAHT	run1	-	-	-	-	-	-	0.3116
SRT	run1	-	0.5954	-	-	-	-	-
	run2	-	0.5943	-	-	-	-	-
	run3	-	0.6045	-	-	-	-	-
Summer	run1	-	-	-	-	-	-	0.4439
	run2	-	-	-	-	-	-	0.4431
	run3	-	-	-	-	-	-	0.4617
Baseline	-	0.3328	0.4042	0.3729	0.4298	0.4757	0.3123	0.2041

Table 5: BLEU scores for "OK" aligned test sentences, into English.

2.2 OCELoT - Biomedical Task

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru	en2zh
aoligei	run1	-	-	-	-	-	-	0.3871
	run2	-	-	-	-	-	-	0.3825
Dtranx	run1	0.3484	0.4918	-	0.4892	0.478	0.3078	0.4114
	run2	-	-	-	0.4752	0.3784	0.1745	0.3625
	run3	-	0.4918	-	0.292	0.2444	-	-
	run4	-	-	0.3473	-	-	-	-
	run5	-	-	0.3518	-	-	-	-
	run6	-	-	0.2384	-	-	-	-
HuaweiTSC	run1	0.3415	-	0.3502	-	-	0.2688	0.4025
	run2	0.3404	-	0.3498	-		0.3059	0.4013
	run3	0.3428	-	0.3556	-	-	0.2673	0.4021
	run4	0.3397	-	0.3613	-	-	-	0.3999
	run5	0.3428	-	-	-	-	-	0.4012
	run6	0.3428	-	-	-	-	-	0.3942
Lan-BridgeMT	run1	0.3110	-	-	-	-	0.2552	0.3786
	run2	0.3167	-	-	-	-	0.2528	0.3690
	run3	-	-	-	-	-	-	0.379
	run3	-	-	-	-	-	-	0.3798
njupt-mtt	run1	0.3394	-	0.3541	-	-	0.2564	0.3653
	run2	-	-	0.3507	-	-	0.2709	0.4025
	run3	-	-	0.3469	-	-	0.2673	0.3987
SPECTRANS	run1	-	-	0.2068	-	-	-	-
	run2	-	-	0.3163	-	-	-	-
	run3	-	-	0.0732	-	-	-	-
	run4	-	-	0.2034	-	-	-	-
ustc-mt	run1	-	-	0.3369	-	-	0.2697	0.4002
	run2	-	-	0.3440	-	-	0.3095	0.3963
	run3	-	-	0.3530	-	-	-	-
	run4	-	-	0.3491	-	-	-	-
	run5	-	-	0.3541	-	-	-	-
	run6	-	-	0.3555	-	-	-	-

Table 6: BLEU scores for ”OK” aligned test sentences, from English.

Teams	**Runs**	**de2en**	**es2en**	**fr2en**	**it2en**	**pt2en**	**ru2en**	**zh2en**
aoligei	run1	-	-	0.4085	-	-	-	-
	run2	-	-	0.3975	-	-	-	-
	run3	-	-	0.4126	-	-	-	-
	run4	-	-	0.4024	-	-	-	-
Dtranx	run1	0.3555	0.5421	0.4494	0.4585	0.5489	0.3840	-
	run2	0.2260	-	0.4638	0.4204	0.5367	-	0.4127
	run3	0.3728	-	0.2691	0.2528	0.2806	0.1934	0.3922
HuaweiTSC	run1	0.3759	-	0.4566	-	-	0.3557	0.4125
	run2	0.3749	-	0.5186	-	-	0.3633	0.4150
	run3	0.3762	-	0.4676	-	-	0.3585	0.4133
	run4	0.3760	-	-	-	-	0.3633	0.4133
	run5	-	-	-	-	-	-	0.4166
	run6	-	-	-	-	-	-	0.4146
Lan-BridgeMT	run1	0.3509	-	-	-	-	0.3171	0.4064
	run2	0.3499	-	-	-	-	0.3124	0.4009
	run3	-	-	-	-	-	-	0.3978
	run4	-	-	-	-	-	-	0.3931
	run5	-	-	-	-	-	-	0.4073
njupt-mtt	run1	0.3709	-	0.4568	-	-	0.3505	0.4139
	run2	-	-	0.4485	-	-	0.3587	0.4132
	run3	-	-	0.4494	-	-	-	-
SPECTRANS	run1	-	-	0.2581	-	-	-	-
	run2	-	-	0.4010	-	-	-	-
	run3	-	-	0.2587	-	-	-	-
	run4	-	-	0.0969	-	-	-	-
szdx	run1	-	-	-	-	-	-	0.3600
ustc-mt	run1	-	-	0.4511	-	-	0.3539	0.4105
	run2	-	-	0.4481	-	-	0.3848	-
	run3	-	-	0.4577	-	-	-	-
	run4	-	-	0.4527	-	-	-	-
	run5	-	-	0.0002	-	-	-	-

Table 7: BLEU scores for ”OK” aligned test sentences, into English.

2.3 OCELoT - General task

Teams	Run	en2de	Run	en2ru	Run	en2zh
AISP-SJTU	-	-	-	-	611	0.377
DLUT	-	-	-	-	651	0.3632
eTranslation	-	-	341	0.2753	-	-
JDExploreAcademy.Vega-MT	843	0.335	509	0.2977	834	0.3982
Lan-Bridge	549	0.3443	556	0.3091	714	0.3797
LanguageX	-	-	-	-	716	0.4157
Manifold	-	-	-	-	336	0.381
ONLINE-A	901	0.3321	912	0.2804	914	0.3794
ONLINE-B	920	0.3488	930	0.309	931	0.4117
ONLINE-W	959	0.3737	966	0.3159	968	0.3942
ONLINE-Y	973	0.3338	983	0.2823	985	0.3779
Online-G	865	0.3376	876	0.2968	878	0.3731
OpenNMT	207	0.3072	-	-	-	-
PROMT	694	0.327	804	0.2918	-	-
SRPOL	-	-	265	0.2724	-	-

Table 8: BLEU scores for "OK" aligned test sentences, from English.

Teams	Run	en2de	Run	en2ru	Run	en2zh
AISP-SJTU	-	-	-	-	648	0.3922
ALMAnaCH-Inria	-	-	710	0,2169	-	-
DLUT	-	-	-	-	654	0.3322
JDExploreAcademy.Vega-MT	809	0.3624	769	0.3785	708	0.4141
Lan-Bridge	587	0.3562	589	0.3886	386	0.4073
LanguageX	-	-	-	-	219	0.4121
LT22	605	0.2469	-	-	-	-
neunlplab	-	-	-	-	847	0.353
ONLINE-A	903	0.3576	913	0.3672	915	0.3667
ONLINE-B	923	0,355	934	0.3827	935	0.4103
ONLINE-W	961	0.3762	967	0.3251	969	0.3741
ONLINE-Y	975	0.3564	984	0.3605	986	0.3689
Online-G	868	0.353	861	0.3769	879	0.3588
PROMT	796	0.3506	70	0.331	-	-
SRPOL	-	-	666	0.3483	-	-

Table 9: BLEU scores for "OK" aligned test sentences, into English.