WMT 2018 - Biomedical task

October 25, 2018

1 Results for Automatic Evaluation

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

* indicates the primary run as informed by the participants.

1.1 MEDLINE dataset

Teams, Runs	de/en	en/de	en/es	en/fr	en/pt	en/ro	es/en	fr/en	pt/en
FOKUS run1						22.17			
FOKUS run2						23.42*			
Hunter MT run1				23.41					
Hunter MT run2				23.24*					
LMU run1	23.93*	18.81*							
LMU run2		18.75							
LMU run3		17.16							
TGF TALP UPC run1							40.49*	25.78*	39.49*
TGF TALP UPC run2							39.06	19.42	38.54
UFRGS run1			39.62*		39.43*		43.31*		42.58*
UFRGS run2			39.77		39.43		43.41		42.58
UHH-DS run1			31.32		34.92	15.40	36.16		41.84
UHH-DS run2			31.05		34.19	15.09	35.17		41.80
UHH-DS run3			31.33*		34.49*	14.77*	36.05*		41.79*

1.2 EDP dataset

Teams, Runs	en/fr
Hunter MT run1	22.20
Hunter MT run2	23.24*

2 Results for Manual Validation

Manual validation using the Appraise tool (3-way ranking task) by comparing translations from either two systems or one system and the reference translation. We only considered one primary run per team as informed by the participants.

2.1 Medline dataset

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

2.2 EDP dataset

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