WMT 2017 - Biomedical task

July 12, 2017

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 Scielo dataset

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*

1.2 EDP dataset

Runs	fr/en	en/fr
baseline	17.47	12.32
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*

1.3 Cochrane dataset

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*	-

1.4 NHS dataset

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*	-

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 Scielo dataset

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

2.2 EDP dataset

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

2.3 Cochrane dataset

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

2.4 NHS dataset

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