

Idx	Rerank	Run	CLEF 2015 Measures			New Measures to Evaluate Understability in Retrieval								
			RBP_r	Res.	uRBP	Res.	RBP_u	Res.	MM_{RBP}	Res.	Unj@10	RBP_r^*	RBP_u^*	MM_{RBP}^*
1		ECNU (1st)	51.57	8.95	50.51	8.95	59.55	10.09	46.22	8.62	0.00	51.57	59.55	46.22
2	No Rerank	KISTI (2nd)	36.72 [◊]	8.06	35.92 [◊]	7.32	64.50	11.54	37.56 [◊]	7.89	0.03	37.07 [◊]	65.31	37.96 [◊]
3		BM25 Baseline	31.20 [◊]	8.76	30.51 [◊]	7.65	67.60	12.20	35.75 [◊]	8.76	0.03	31.57 [◊]	68.94 [◊]	36.42 [◊]
4	SMOG Top 15	Based on ECNU	38.16 ^{†◊}	20.09	37.36 ^{†◊}	8.95	55.11 ^{†◊}	22.14	37.59 ^{†◊}	18.98	0.14	45.45 ^{†◊}	62.58	43.65
5		Based on KISTI	31.28 ^{†◊}	10.21	30.59 ^{†◊}	7.48	67.05	13.95	34.23 [◊]	10.05	0.05	33.12 [◊]	69.10 [◊]	36.02 [◊]
6		Based on BM25	24.39 ^{†◊}	11.46	23.83 ^{†◊}	7.16	67.69	17.66	28.97 ^{†◊}	11.46	0.10	26.28 ^{†◊}	72.82 ^{†◊}	31.53 ^{†◊}
7	SMOG Top 20	Based on ECNU	34.88 ^{†◊}	25.96	34.13 ^{†◊}	8.95	54.15 ^{†◊}	27.91	35.34 ^{†◊}	25.02	0.20	44.87 ^{†◊}	66.20 ^{†◊}	44.88
8		Based on KISTI	28.01 ^{†◊}	11.05	27.39 ^{†◊}	7.32	66.61	15.44	30.93 ^{†◊}	10.82	0.08	30.87 ^{†◊}	69.99 ^{†◊}	33.81 ^{†◊}
9		Based on BM25	22.83 ^{†◊}	13.48	22.31 ^{†◊}	7.32	63.61	22.44	26.52 ^{†◊}	13.48	0.15	26.22 ^{†◊}	72.72 ^{†◊}	30.87 ^{†◊}
10	SMOG Top 50	Based on ECNU	21.83 ^{†◊}	36.63	21.34 ^{†◊}	8.95	41.71 ^{†◊}	46.89	23.65 ^{†◊}	36.63	0.45	39.88 ^{†◊}	71.61 ^{†◊}	43.86
11		Based on KISTI	21.00 ^{†◊}	18.67	20.53 ^{†◊}	7.32	59.72	26.67	24.88 ^{†◊}	18.67	0.23	27.20 ^{†◊}	72.47 ^{†◊}	31.90 ^{†◊}
12		Based on BM25	15.20 ^{†◊}	18.73	14.86 ^{†◊}	6.51	49.87 ^{†◊}	36.41	17.79 ^{†◊}	18.73	0.32	21.22 ^{†◊}	73.17 ^{†◊}	25.78 ^{†◊}
13	XGB Top 15	Based on ECNU	39.77 ^{†◊}	21.59	38.90 ^{†◊}	9.11	54.84 ^{†◊}	24.22	38.35 ^{†◊}	20.72	0.15	47.00 ^{†◊}	64.28 ^{†◊}	45.22
14		Based on KISTI	31.36 ^{†◊}	8.95	30.63 ^{†◊}	6.83	68.63 [◊]	13.63	33.78 [◊]	8.78	0.05	32.80 ^{†◊}	70.87 ^{†◊}	35.24 [◊]
15		Based on BM25	23.38 ^{†◊}	11.36	22.84 ^{†◊}	7.16	66.79	19.43	27.60 ^{†◊}	11.36	0.10	26.35 ^{†◊}	73.48 ^{†◊}	31.35 ^{†◊}
16	XGB Top 20	Based on ECNU	34.91 ^{†◊}	27.19	34.12 ^{†◊}	9.11	52.69 ^{†◊}	30.19	34.89 ^{†◊}	25.95	0.25	46.71 ^{†◊}	67.26 ^{†◊}	45.81
17		Based on KISTI	29.04 ^{†◊}	10.10	28.35 ^{†◊}	6.51	69.10 [◊]	15.86	32.00 ^{†◊}	9.94	0.07	31.82 [◊]	73.19 ^{†◊}	34.84 [◊]
18		Based on BM25	21.83 ^{†◊}	12.88	21.32 ^{†◊}	6.51	64.16	25.29	25.47 ^{†◊}	12.88	0.16	26.56 ^{†◊}	76.26 ^{†◊}	31.45 ^{†◊}
19	XGB Top 50	Based on ECNU	22.75 ^{†◊}	33.65	22.22 ^{†◊}	8.79	43.36 ^{†◊}	49.18	24.90 ^{†◊}	33.65	0.45	40.69 ^{†◊}	71.39 ^{†◊}	44.65
20		Based on KISTI	18.69 ^{†◊}	17.21	18.25 ^{†◊}	6.67	56.87 [†]	33.37	21.82 ^{†◊}	17.21	0.29	27.69 ^{†◊}	76.44 ^{†◊}	32.53 [◊]
21		Based on BM25	17.47 ^{†◊}	20.95	17.00 ^{†◊}	6.02	47.21 ^{†◊}	43.81	19.22 ^{†◊}	20.98	0.41	27.09 ^{†◊}	79.77 ^{†◊}	31.35 [◊]
22	RRF (XGB & Orig.) Top 15	Based on ECNU	47.81 ^{†◊}	12.85	46.78 ^{†◊}	8.95	60.04	15.04	44.69	12.34	0.10	50.09	62.98 ^{†◊}	46.85
23		Based on KISTI	33.78 ^{†◊}	7.71	33.02 ^{†◊}	6.83	68.57 ^{†◊}	12.02	35.72 [†]	7.55	0.03	34.34 ^{†◊}	69.67 ^{†◊}	36.26 [†]
24		Based on BM25	26.85 ^{†◊}	12.13	26.23 ^{†◊}	7.48	66.64	15.98	31.52 ^{†◊}	12.13	0.07	28.35 ^{†◊}	70.65 ^{†◊}	33.70 ^{†◊}
25	RRF (XGB & Orig.) Top 20	Based on ECNU	46.49 ^{†◊}	15.16	45.48 ^{†◊}	9.11	59.95	17.12	43.69	14.54	0.12	49.95	64.30 ^{†◊}	47.01
26		Based on KISTI	32.72 ^{†◊}	8.45	31.97 ^{†◊}	7.00	69.06 ^{†◊}	12.77	35.14 [◊]	8.28	0.04	33.83 ^{†◊}	70.83 ^{†◊}	36.29 ^{†◊}
27		Based on BM25	25.64 ^{†◊}	12.51	25.05 ^{†◊}	7.16	66.55	17.31	30.14 ^{†◊}	12.51	0.09	27.64 ^{†◊}	71.79 ^{†◊}	32.96 ^{†◊}
28	RRF (XGB & Orig.) Top 50	Based on ECNU	38.97 ^{†◊}	21.40	38.08 ^{†◊}	8.79	57.37	25.44	39.54 ^{†◊}	19.83	0.24	47.00 ^{†◊}	67.95 ^{†◊}	47.12
29		Based on KISTI	27.78 ^{†◊}	11.24	27.13 ^{†◊}	6.83	67.83	16.75	31.07 ^{†◊}	11.24	0.09	31.33 ^{†◊}	73.06 ^{†◊}	34.85 ^{†◊}
30		Based on BM25	19.28 ^{†◊}	17.11	18.86 ^{†◊}	7.00	57.40 [†]	27.02	22.78 ^{†◊}	17.11	0.19	25.06 ^{†◊}	71.56 ^{†◊}	30.26 ^{†◊}
31	RRF (SMOG & Orig.) Top 15	Based on ECNU	47.23 ^{†◊}	12.41	46.24 ^{†◊}	8.95	59.57	13.88	43.96 ^{†◊}	11.80	0.08	49.21 ^{†◊}	61.45	45.42
32		Based on KISTI	34.84 [◊]	8.34	34.08 [◊]	7.48	67.87 ^{†◊}	11.84	36.86 [◊]	8.18	0.03	35.32 [◊]	68.84 ^{†◊}	37.37 [◊]
33		Based on BM25	27.16 ^{†◊}	10.79	26.55 ^{†◊}	7.32	66.44	14.65	32.24 ^{†◊}	10.79	0.05	28.21 ^{†◊}	69.38 ^{†◊}	33.67 ^{†◊}
34	RRF (SMOG & Orig.) Top 20	Based on ECNU	45.91 ^{†◊}	14.14	44.94 ^{†◊}	9.11	59.46	15.28	43.50 ^{†◊}	13.55	0.09	48.67 ^{†◊}	62.46 ^{†◊}	45.93
35		Based on KISTI	34.17 [◊]	8.54	33.42 [◊]	7.48	68.23 ^{†◊}	12.09	36.23 [◊]	8.38	0.03	34.79 [◊]	69.30 ^{†◊}	36.89 [◊]
36		Based on BM25	25.80 ^{†◊}	12.33	25.22 ^{†◊}	7.48	64.98	16.18	30.69 ^{†◊}	12.33	0.08	27.24 ^{†◊}	69.14 ^{†◊}	32.68 ^{†◊}
37	RRF (SMOG & Orig.) Top 50	Based on ECNU	39.06 ^{†◊}	20.77	38.20 ^{†◊}	9.27	57.68	23.16	39.55 ^{†◊}	20.28	0.18	45.59 ^{†◊}	65.82 ^{†◊}	46.13
38		Based on KISTI	29.51 ^{†◊}	10.35	28.83 ^{†◊}	7.48	68.60 ^{†◊}	14.71	33.37	10.35	0.08	31.48 ^{†◊}	71.70 ^{†◊}	35.54
39		Based on BM25	18.70 ^{†◊}	17.12	18.29 ^{†◊}	7.16	57.50 [†]	24.59	22.73 ^{†◊}	17.12	0.15	22.95 ^{†◊}	68.61	28.21 ^{†◊}
40	XGB LeToR	LTR 1 on BM25	24.86 ^{†◊}	17.39	24.32 ^{†◊}	7.81	55.60 ^{†◊}	24.11	28.89 ^{†◊}	17.39	0.22	29.67 [◊]	66.41	34.76 [◊]
41		LTR 2 on BM25	30.72 [◊]	21.25	30.08 [◊]	8.46	48.87 ^{†◊}	28.82	31.76 [◊]	18.99	0.26	37.09 ^{†◊}	61.89 [†]	39.17 [◊]
42		LTR 3 on BM25	28.92 [◊]	24.35	28.32 [◊]	8.46	49.02 ^{†◊}	32.11	30.14 ^{†◊}	23.83	0.31	37.32 [◊]	63.86 [†]	39.84 [◊]
43		LTR 4 on BM25	25.65 ^{†◊}	25.72	25.09 [†]	8.30	49.00 ^{†◊}	33.39	27.45 ^{†◊}	24.40	0.33	35.82 [◊]	66.14	38.21 [◊]
44		LTR 5 on BM25	30.21 [◊]	20.79	29.59 [◊]	8.62	48.47 ^{†◊}	27.88	30.95 ^{†◊}	19.99	0.25	37.11 [◊]	61.25 ^{†◊}	39.15 [◊]

Table 1. Results obtained by integrating understandability estimations within retrieval methods on CLEF 2015. Baseline runs are reported at table indices 1–3 (the index column is labeled Idx). Reranking experiments are reported at indices 4–21. Fusion experiments are reported at indices 22–39. Learning to rank experiments are reported at indices 40–44. All measures were calculated up to rank $n = 10$. The highest result of each set of experiments is reported in boldface for each measure. Statistically significant differences compared to ECNU are indicated with \diamond , while differences between an original run (indices 1-3) and its modifications are indicated with \dagger (paired, two-tail t-test, $P < 0.05$).

Idx	Rerank	Run	CLEF 2016 Measures				New Measures to Evaluate Understandability in Retrieval							
			RBP_r	Res.	uRBP	Res.	RBP_u	Res.	MM_{RBP}	Res.	Unj@10	RBP_r^*	RBP_u^*	MM_{RBP}^*
1		GUIR (1st)	28.11	7.65	18.12	7.19	45.69	8.86	25.61	6.50	0.01	28.29	46.03	25.79
2	No Rerank	ECNU (2nd)	27.70	7.37	17.55	7.34	43.89 [◊]	8.66	25.35	6.26	0.01	27.77	44.18 [◊]	25.48
3		BM25 Baseline	25.28 [◊]	8.24	16.05 [◊]	6.94	42.08 [◊]	10.97	22.97 [◊]	7.19	0.06	26.01 [◊]	43.89 [◊]	23.93 [◊]
4	Dale-Chall Top 15	Based on GUIR	24.70 ^{†◊}	8.70	16.83 ^{†◊}	7.27	49.10 ^{†◊}	10.62	24.94	7.50	0.03	25.24 ^{†◊}	50.33 ^{†◊}	25.54
5		Based on ECNU	24.78 ^{†◊}	7.83	16.64 [◊]	7.16	48.88 ^{†◊}	9.71	24.80	6.50	0.02	25.12 ^{†◊}	49.64 ^{†◊}	25.21
6		Based on BM25	23.22 ^{†◊}	8.78	15.85 [◊]	6.94	47.09 ^{†◊}	11.83	24.01	7.42	0.07	24.04 ^{†◊}	48.60 ^{†◊}	24.82
7	Dale-Chall Top 20	Based on GUIR	22.19 ^{†◊}	9.37	15.36 ^{†◊}	6.98	48.71 ^{†◊}	12.30	23.21 ^{†◊}	8.12	0.06	23.26 ^{†◊}	51.39 ^{†◊}	24.45 ^{†◊}
8		Based on ECNU	23.01 ^{†◊}	8.93	15.70 ^{†◊}	6.91	48.99 ^{†◊}	11.69	23.73 ^{†◊}	7.80	0.05	23.84 ^{†◊}	51.00 ^{†◊}	24.66
9		Based on BM25	21.58 ^{†◊}	9.51	14.83 ^{†◊}	7.02	46.99 [†]	13.00	22.89 [◊]	8.06	0.09	22.93 ^{†◊}	49.55 ^{†◊}	24.26
10	Dale-Chall Top 50	Based on GUIR	16.18 ^{†◊}	15.24	11.56 ^{†◊}	6.80	41.79 ^{†◊}	24.49	18.10 ^{†◊}	14.42	0.22	20.90 ^{†◊}	53.28 ^{†◊}	23.27 ^{†◊}
11		Based on ECNU	16.88 ^{†◊}	17.37	11.78 ^{†◊}	7.30	40.76 ^{†◊}	23.77	18.30 ^{†◊}	15.57	0.24	21.34 ^{†◊}	52.07 ^{†◊}	23.33 ^{†◊}
12		Based on BM25	15.06 ^{†◊}	15.35 ^{†◊}	10.55	6.62	40.03 [◊]	23.88	16.55 ^{†◊}	13.83	0.24	19.42 ^{†◊}	51.69 ^{†◊}	21.59 ^{†◊}
13	XGB Top 15	Based on GUIR	25.16 ^{†◊}	8.09	17.27 ^{†◊}	7.12	50.96 ^{†◊}	10.11	25.16	6.89	0.02	25.61 ^{†◊}	52.00 ^{†◊}	25.68
14		Based on ECNU	24.18 ^{†◊}	7.69	16.54 [◊]	7.09	50.00 ^{†◊}	9.91	24.56	6.65	0.02	24.56 ^{†◊}	50.74 ^{†◊}	25.01
15		Based on BM25	22.33 ^{†◊}	8.14	15.46	6.76	47.90 ^{†◊}	12.13	22.89 [◊]	7.25	0.07	23.11 ^{†◊}	49.43 ^{†◊}	23.69 [◊]
16	XGB Top 20	Based on GUIR	22.38 ^{†◊}	9.49	15.61 ^{†◊}	7.05	50.45 ^{†◊}	12.08	23.30 ^{†◊}	8.16	0.05	23.62 ^{†◊}	52.98 ^{†◊}	24.68
17		Based on ECNU	22.95 ^{†◊}	8.82	15.95 ^{†◊}	7.02	50.42 ^{†◊}	11.70	23.97 [◊]	7.56	0.04	23.68 ^{†◊}	52.15 ^{†◊}	24.73
18		Based on BM25	20.65 ^{†◊}	9.42	14.46 ^{†◊}	6.84	47.74 ^{†◊}	13.56	21.93 [◊]	8.34	0.09	21.98 ^{†◊}	50.28 ^{†◊}	23.27 [◊]
19	XGB Top 50	Based on GUIR	16.65 ^{†◊}	15.73	12.39 ^{†◊}	6.84	43.49 ^{†◊}	23.63	18.70 ^{†◊}	13.74	0.22	21.13 ^{†◊}	55.07 ^{†◊}	23.58 ^{†◊}
20		Based on ECNU	16.19 ^{†◊}	17.01	11.82 ^{†◊}	7.27	43.05 [◊]	24.75	18.27 ^{†◊}	14.41	0.24	20.16 ^{†◊}	54.70 ^{†◊}	22.96 ^{†◊}
21		Based on BM25	15.43 ^{†◊}	15.37	11.33 ^{†◊}	6.48	41.93 ^{†◊}	23.65	17.43 ^{†◊}	13.40	0.26	19.58 ^{†◊}	54.04 ^{†◊}	22.17 ^{†◊}
22	RRF (XGB & Orig.) Top 15	Based on GUIR	27.23 ^{†◊}	7.76	18.31	7.23	49.69 ^{†◊}	9.18	26.49 ^{†◊}	6.62	0.01	27.46 ^{†◊}	50.07 ^{†◊}	26.69 ^{†◊}
23		Based on ECNU	26.60 ^{†◊}	7.41	17.81	7.19	48.67 ^{†◊}	8.80	26.02	6.09	0.01	26.76 ^{†◊}	49.10 ^{†◊}	26.27 ^{†◊}
24		Based on BM25	24.57 [◊]	8.15	16.51 [◊]	6.91	46.76 [†]	11.23	24.16 [†]	7.20	0.06	25.32 [◊]	48.52 [†]	25.08 [†]
25	RRF (XGB & Orig.) Top 20	Based on GUIR	26.21 ^{†◊}	7.96	17.73	7.19	50.29 ^{†◊}	9.58	25.89	6.73	0.03	26.53 ^{†◊}	50.98 ^{†◊}	26.25
26		Based on ECNU	26.15 ^{†◊}	7.64	17.69	7.09	49.70 ^{†◊}	9.28	26.07	6.39	0.02	26.38 ^{†◊}	50.32 ^{†◊}	26.35
27		Based on BM25	24.04 ^{†◊}	8.24	16.32 [◊]	6.87	47.69 ^{†◊}	11.40	24.08 ^{†◊}	7.35	0.06	24.82 ^{†◊}	49.52 ^{†◊}	25.01 [†]
28	RRF (XGB & Orig.) Top 50	Based on GUIR	24.09 ^{†◊}	9.44	16.85 ^{†◊}	7.02	50.55 ^{†◊}	11.76	24.76	8.01	0.07	25.08 ^{†◊}	52.84 ^{†◊}	25.84
29		Based on ECNU	24.17 ^{†◊}	8.67	16.75 [◊]	7.12	50.63 ^{†◊}	11.66	25.00	7.61	0.07	24.90 ^{†◊}	52.50 ^{†◊}	25.84
30		Based on BM25	22.28 ^{†◊}	8.87	15.50	6.76	48.79 ^{†◊}	12.90	23.13 ^{†◊}	7.82	0.10	23.46 ^{†◊}	51.89 ^{†◊}	24.57
31	RRF (Dale-Chall & Orig.) Top 15	Based on GUIR	26.79 ^{†◊}	8.06	17.94	7.37	48.43 ^{†◊}	9.34	26.16	6.81	0.02	27.06 ^{†◊}	48.98 ^{†◊}	26.46
32		Based on ECNU	26.81	7.58	17.70	7.30	47.98 ^{†◊}	9.04	26.25	6.30	0.02	27.00	48.53 ^{†◊}	26.50 [†]
33		Based on BM25	24.91 [◊]	8.12	16.58 [◊]	7.09	45.85 [†]	10.75	24.67 [†]	6.99	0.06	25.68 [◊]	47.55 ^{†◊}	25.56 [†]
34	RRF (Dale-Chall & Orig.) Top 20	Based on GUIR	25.91 ^{†◊}	8.20	17.49 ^{†◊}	7.27	48.76 ^{†◊}	9.92	25.75	6.93	0.03	26.31 ^{†◊}	49.66 ^{†◊}	26.19
35		Based on ECNU	26.38 ^{†◊}	8.03	17.50	7.44	48.55 ^{†◊}	9.50	26.25	6.72	0.03	26.64 ^{†◊}	49.37 ^{†◊}	26.57 [◊]
36		Based on BM25	24.73 [◊]	8.21	16.63 [◊]	7.12	46.63 [†]	11.00	24.85 [†]	7.11	0.06	25.51 [◊]	48.44 ^{†◊}	25.77 [†]
37	RRF (Dale-Chall & Orig.) Top 50	Based on GUIR	23.29 ^{†◊}	9.76	16.15 ^{†◊}	7.12	49.04 ^{†◊}	12.47	24.13	8.37	0.07	24.48 ^{†◊}	51.45 ^{†◊}	25.41
38		Based on ECNU	24.16 ^{†◊}	9.12	16.37 [◊]	7.34	48.84 ^{†◊}	11.64	24.78	7.86	0.07	25.06 ^{†◊}	50.88 ^{†◊}	25.80
39		Based on BM25	22.64 ^{†◊}	9.26	15.47 [◊]	7.16	47.25 ^{†◊}	12.79	23.67	7.86	0.09	23.89 ^{†◊}	50.26 ^{†◊}	25.11
40	XGB LeToR	LTR 1 on BM25	20.42 ^{†◊}	17.61	13.00 ^{†◊}	7.41	32.17 ^{†◊}	24.61	18.39 ^{†◊}	14.41	0.28	25.25 [◊]	43.19 [◊]	23.83 [◊]
41		LTR 2 on BM25	24.98 ^{†◊}	19.83	15.30 ^{†◊}	8.09	35.09 ^{†◊}	25.14	22.26 [◊]	17.50	0.24	30.41	46.09	28.28 ^{†◊}
42		LTR 3 on BM25	26.35 [†]	20.48	15.88 ^{†◊}	8.16	34.73 ^{†◊}	24.69	21.81 [†]	17.41	0.22	32.25 [◊]	45.44	28.22 ^{†◊}
43		LTR 4 on BM25	16.16 ^{†◊}	19.48	10.76 ^{†◊}	7.27	36.75 ^{†◊}	28.51	16.77 ^{†◊}	17.80	0.29	22.20 ^{†◊}	50.06 ^{†◊}	23.32 [◊]
44		LTR 5 on BM25	26.76 [†]	20.48	16.19 [◊]	8.34	35.26 ^{†◊}	24.13	22.96	17.59	0.22	32.60 [†]	45.87	29.20 ^{†◊}

Table 2. Results obtained by integrating understandability estimations within retrieval methods on CLEF 2016. Baseline runs are reported at table indices 1–3 (the index column is labeled Idx). Reranking experiments are reported at indices 4–21. Fusion experiments are reported at indices 22–39. Learning to rank experiments are reported at indices 40–44. All measures were calculated up to rank $n = 10$. The highest result of each set of experiments is reported in boldface for each measure. Statistically significant differences compared to ECNU are indicated with \diamond , while differences between an original run (indices 1-3) and its modifications are indicated with \dagger (paired, two-tail t-test, $P < 0.05$).