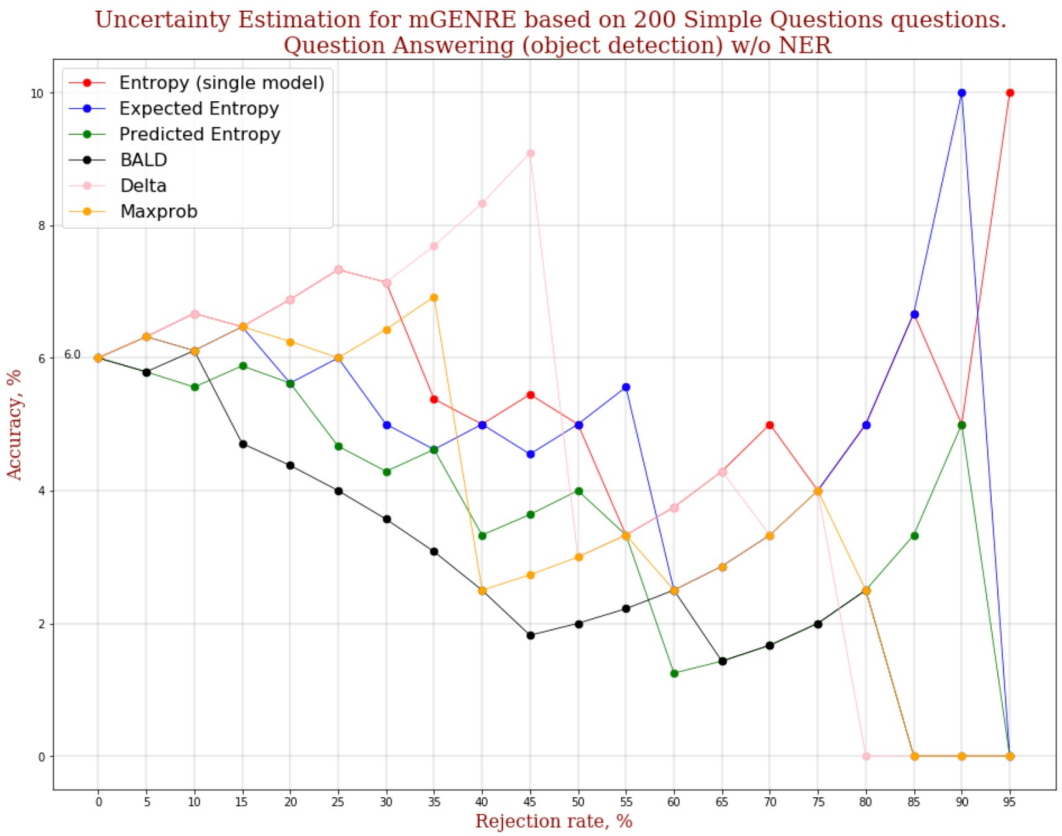
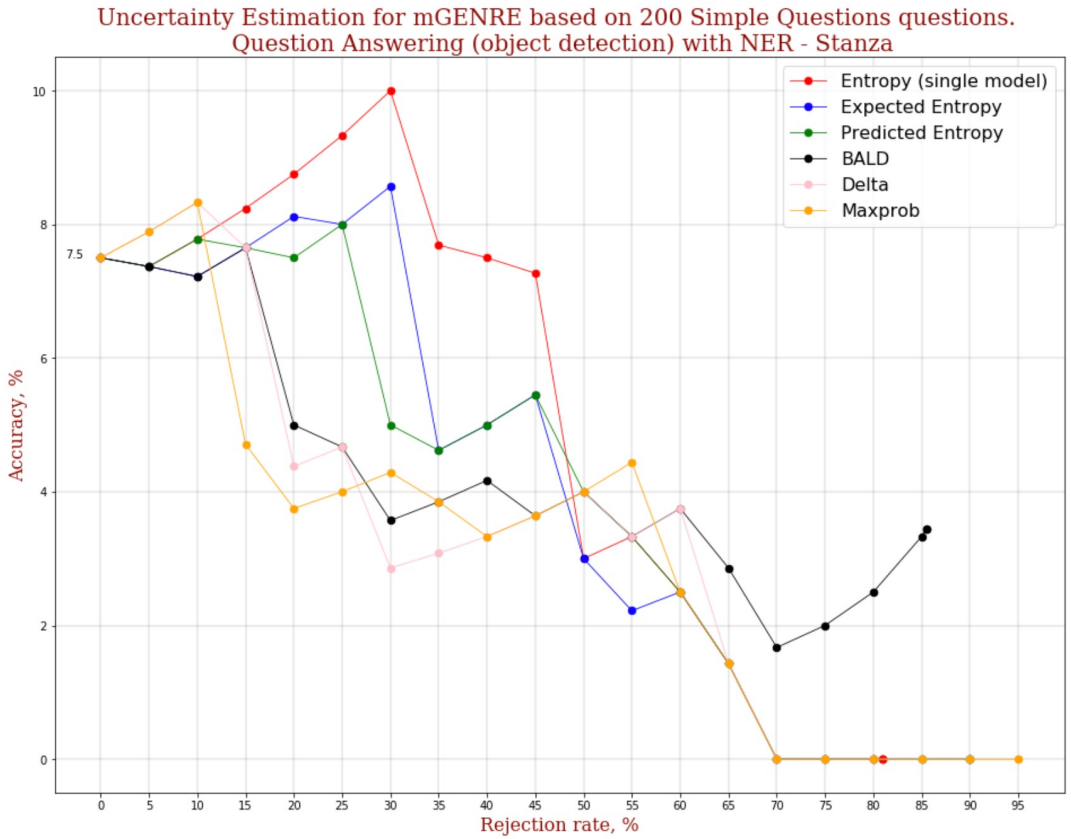
mGenre
Uncertainty Estimation

Experiments conducted on 3 datasets:

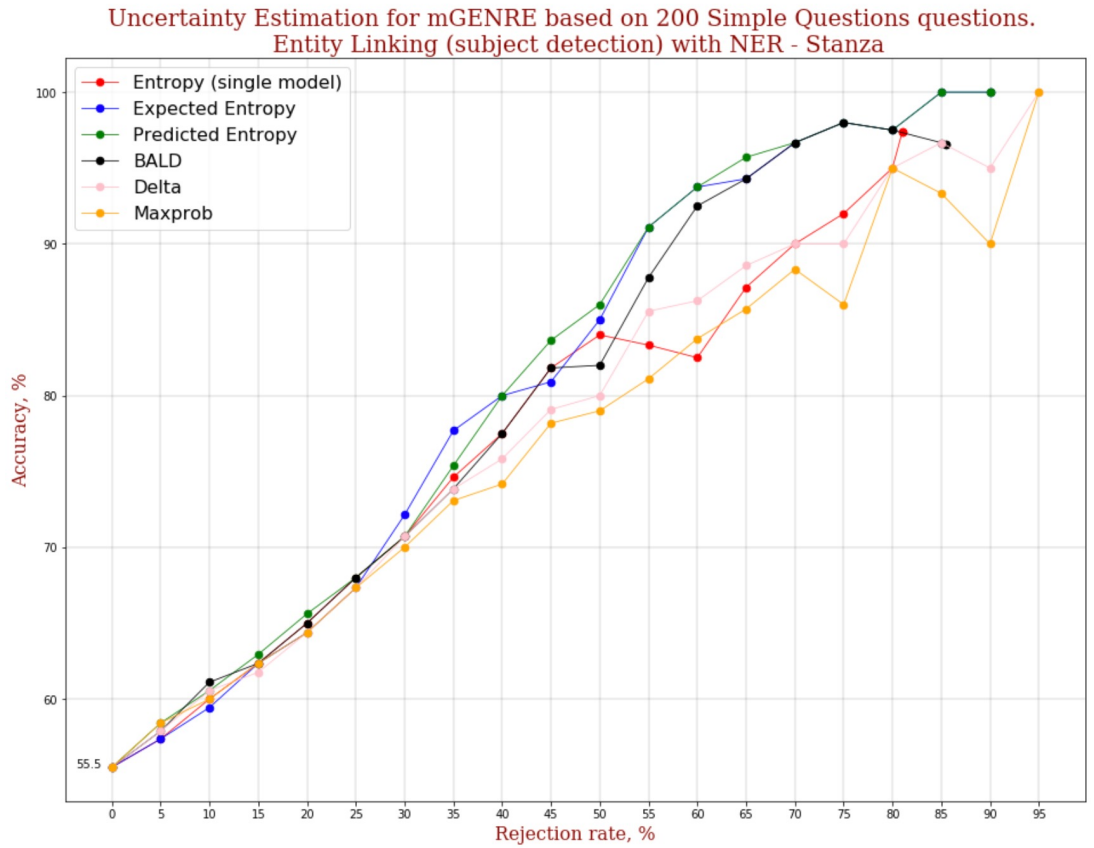
- Simple Questions (both with and without NER)
  - English
  - English (questions only with 1 answer)
- RUBQ 2.0 (both with and without NER)
  - English
  - Russian
- Mewsli-9 (marked where only without NER)
  - English
  - Deutsch
  - Spanish
  - Japanese
  - Persian
  - Serbian (without NER)
  - Turkish
  - Javanese (without NER)

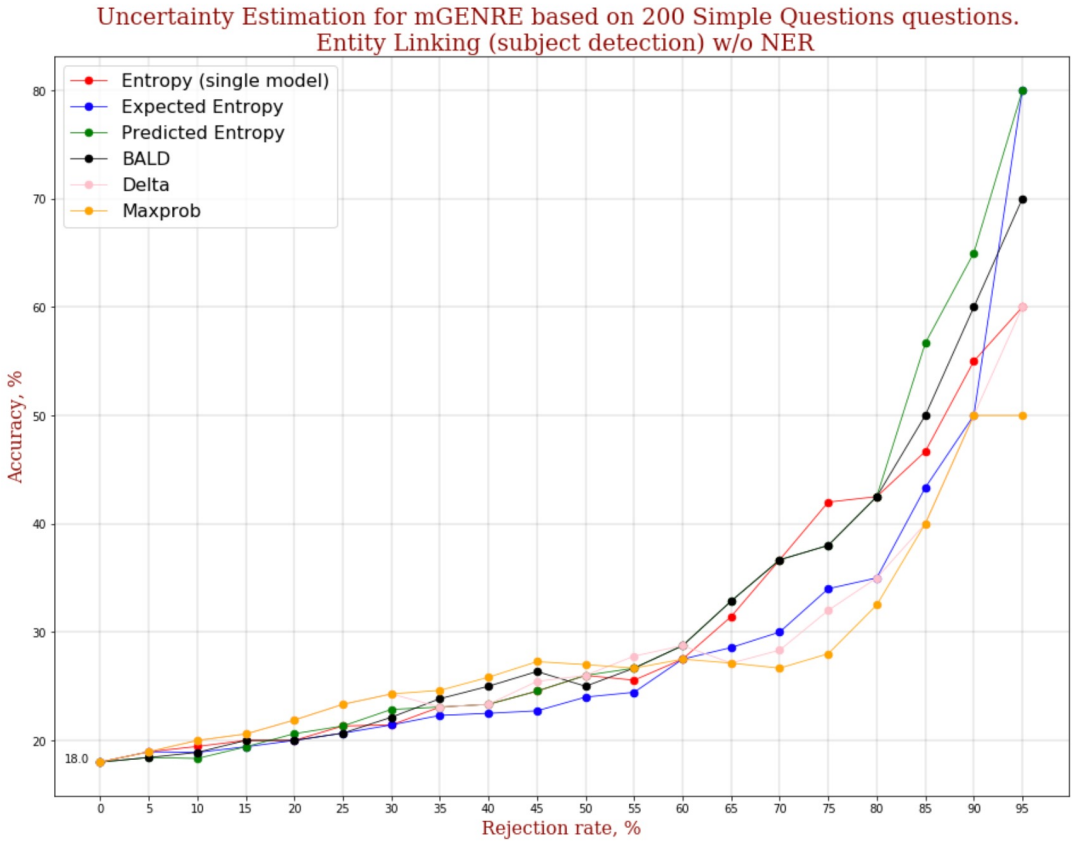
Part 1. Simple Questions



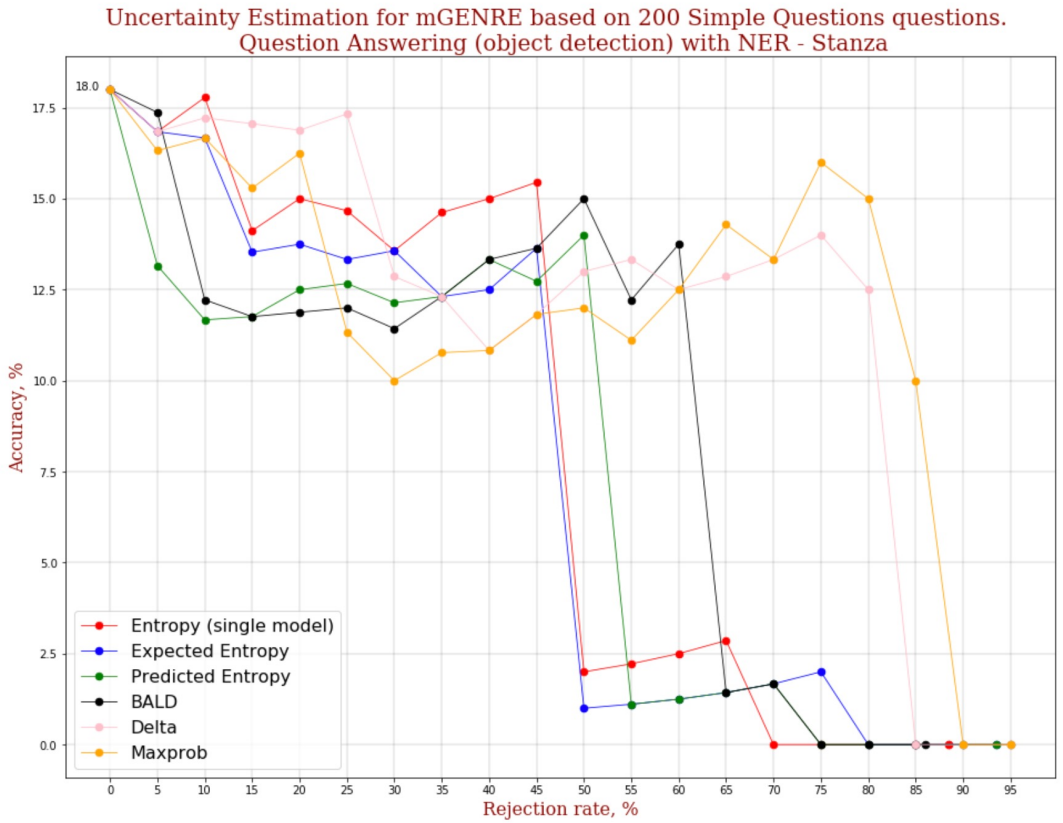


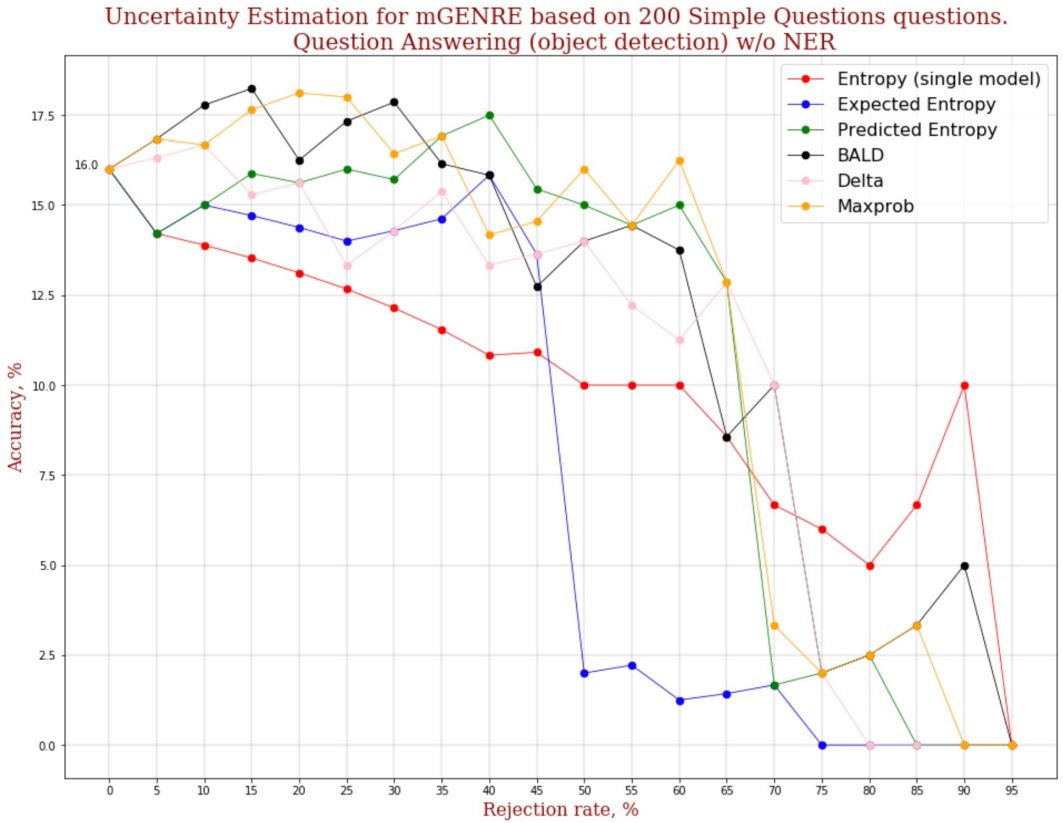
Simple Questions (1.1 “Full”)
Question Answering. Object detection



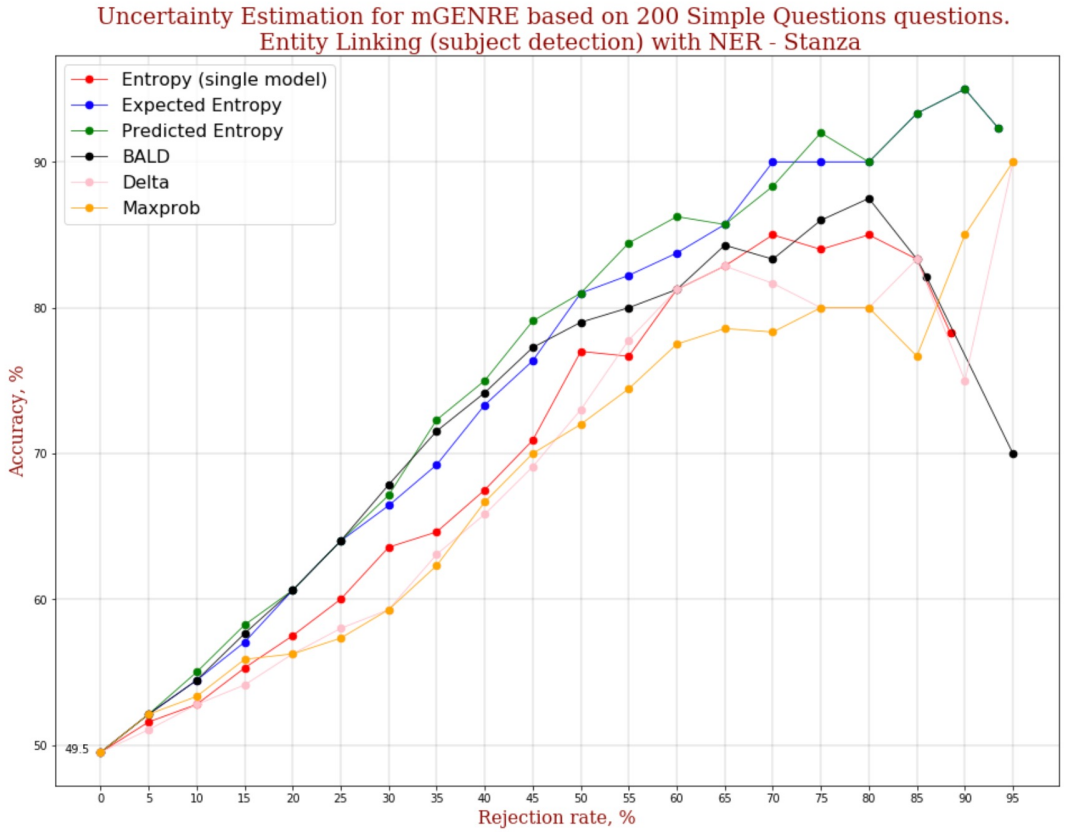


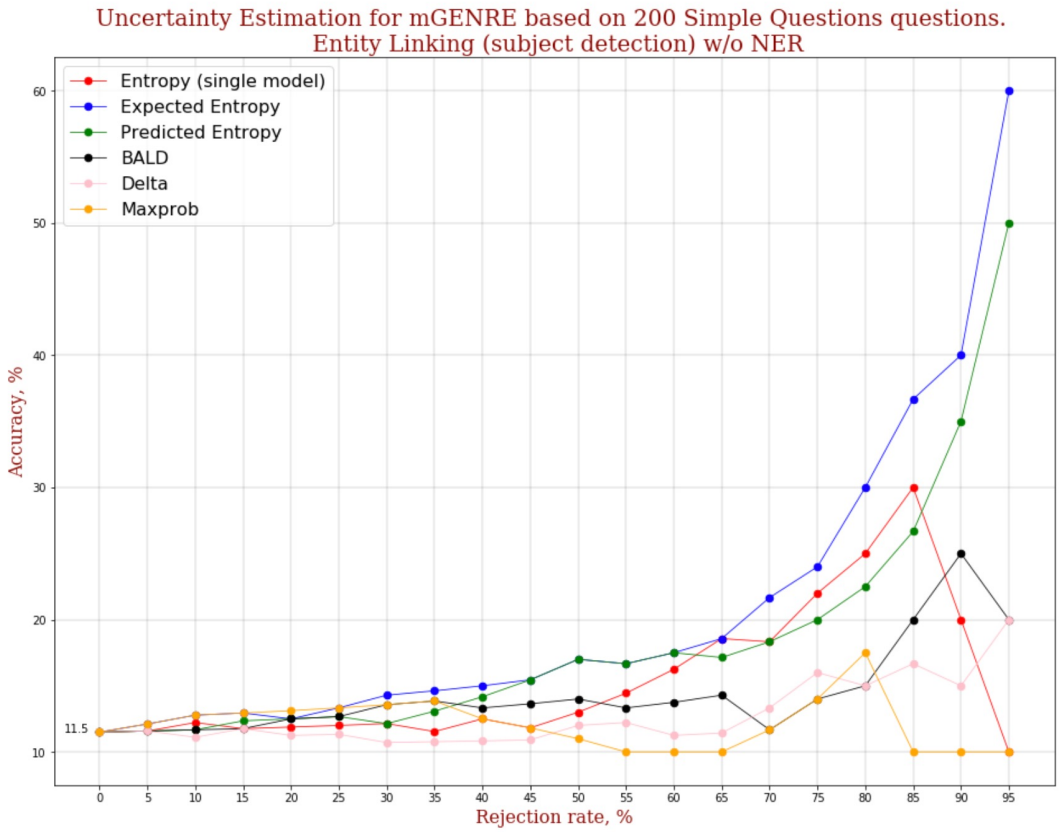
Simple Questions (1.2 “Full”)
Entity Linking. Subject detection





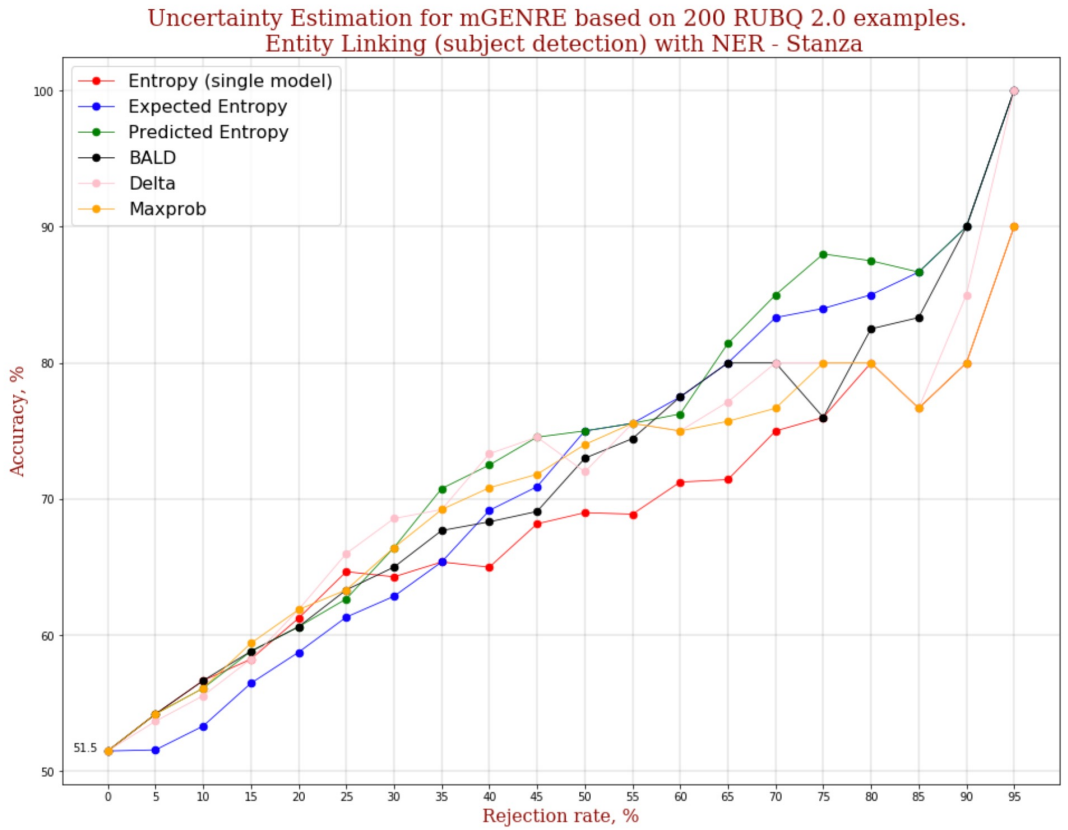
Simple Questions (2.1 “One answer”) Question answering. Object detection

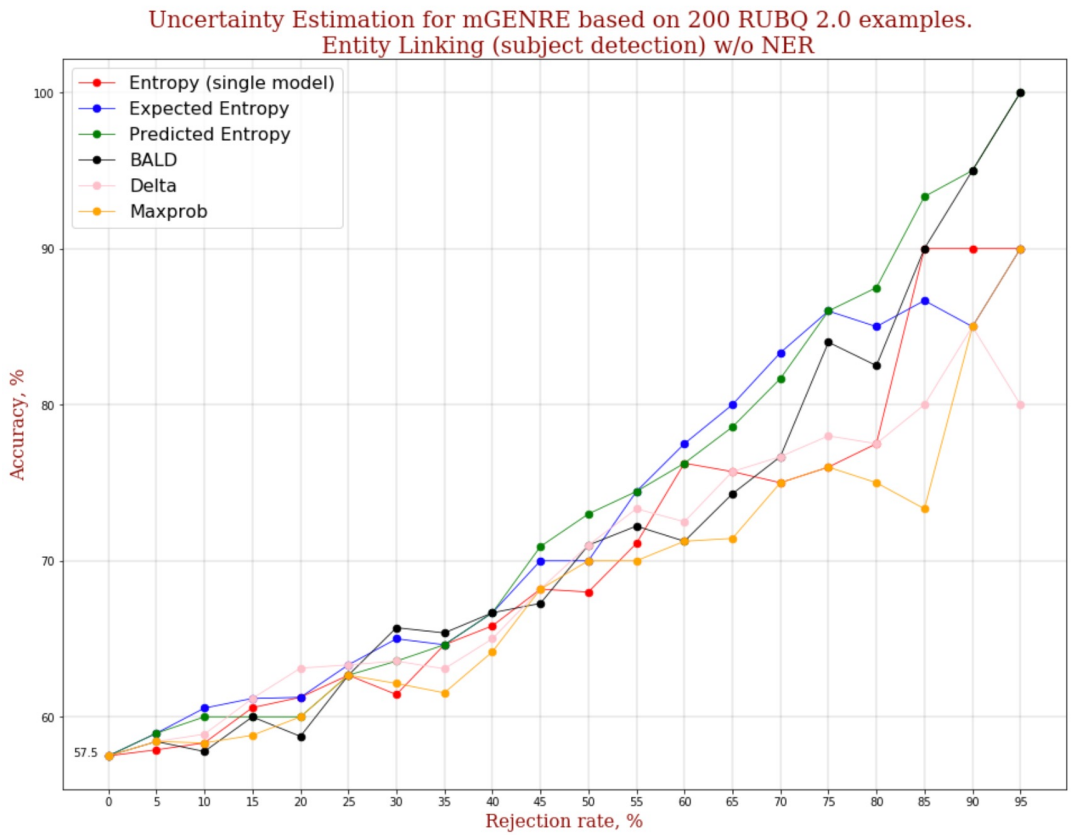




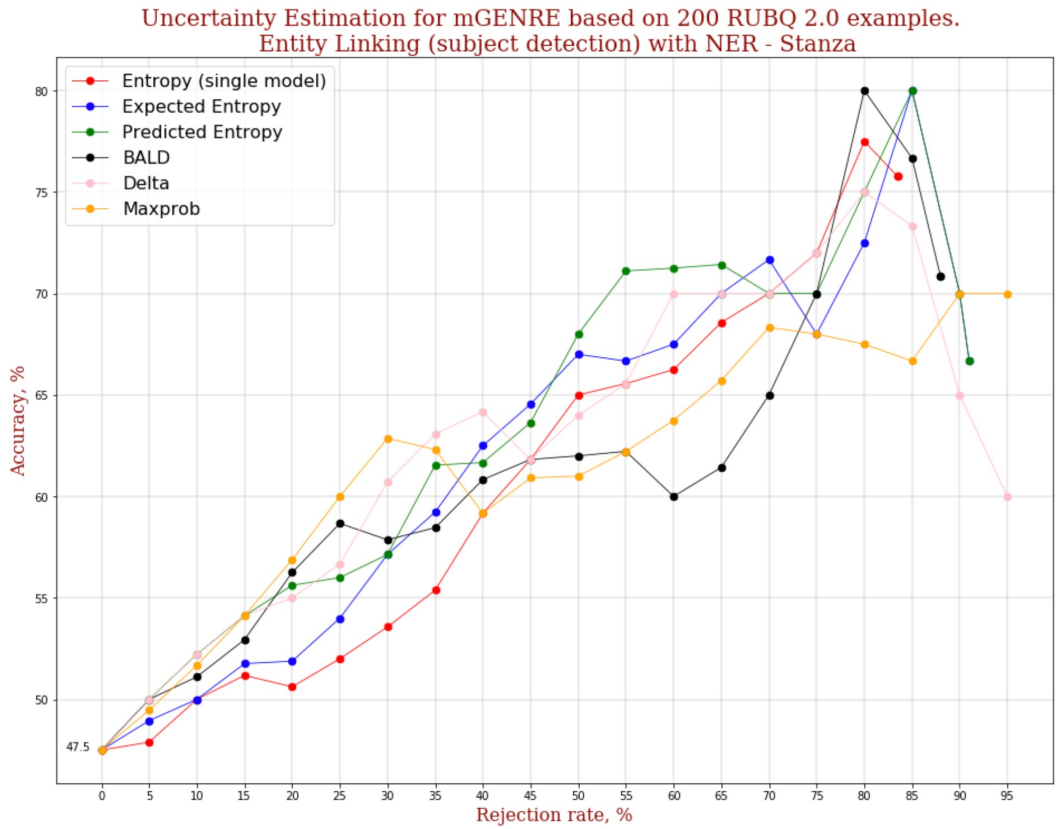
Simple Questions (2.2 “One answer”) Entity Linking. Subject detection

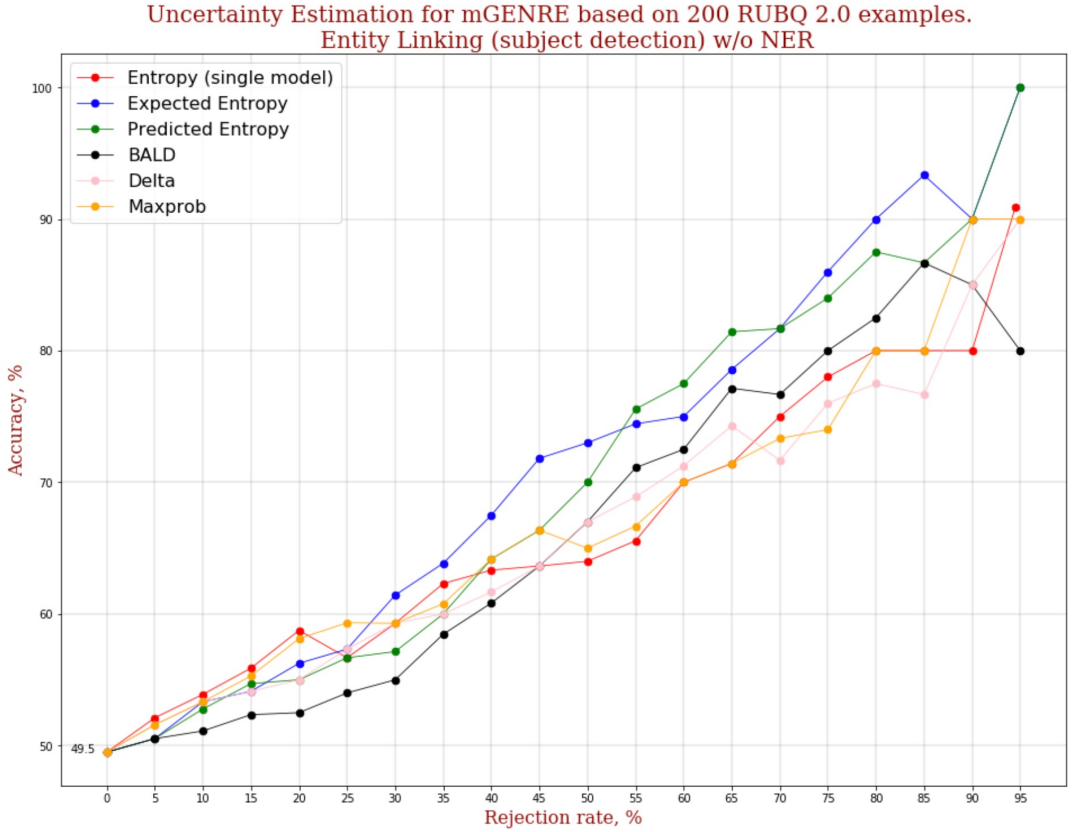
Part 2. RUBQ 2.0





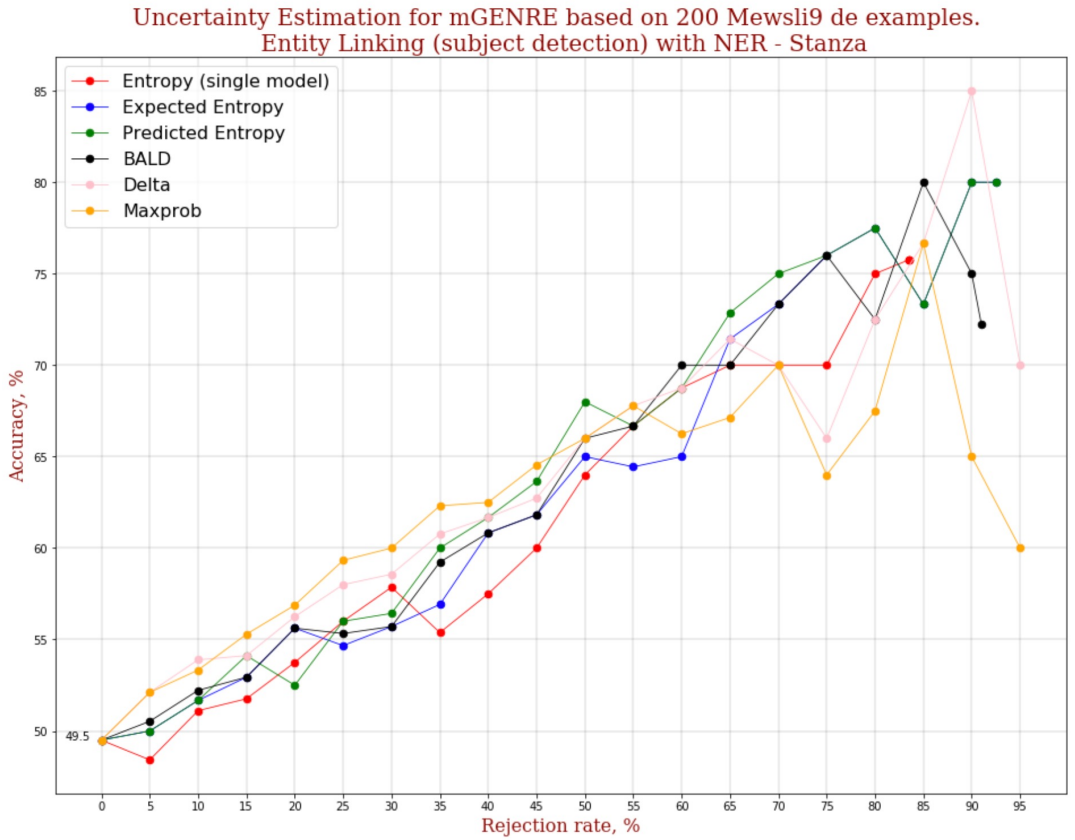
RUBQ 2.0 (Russian)
Entity Linking. Subject detection

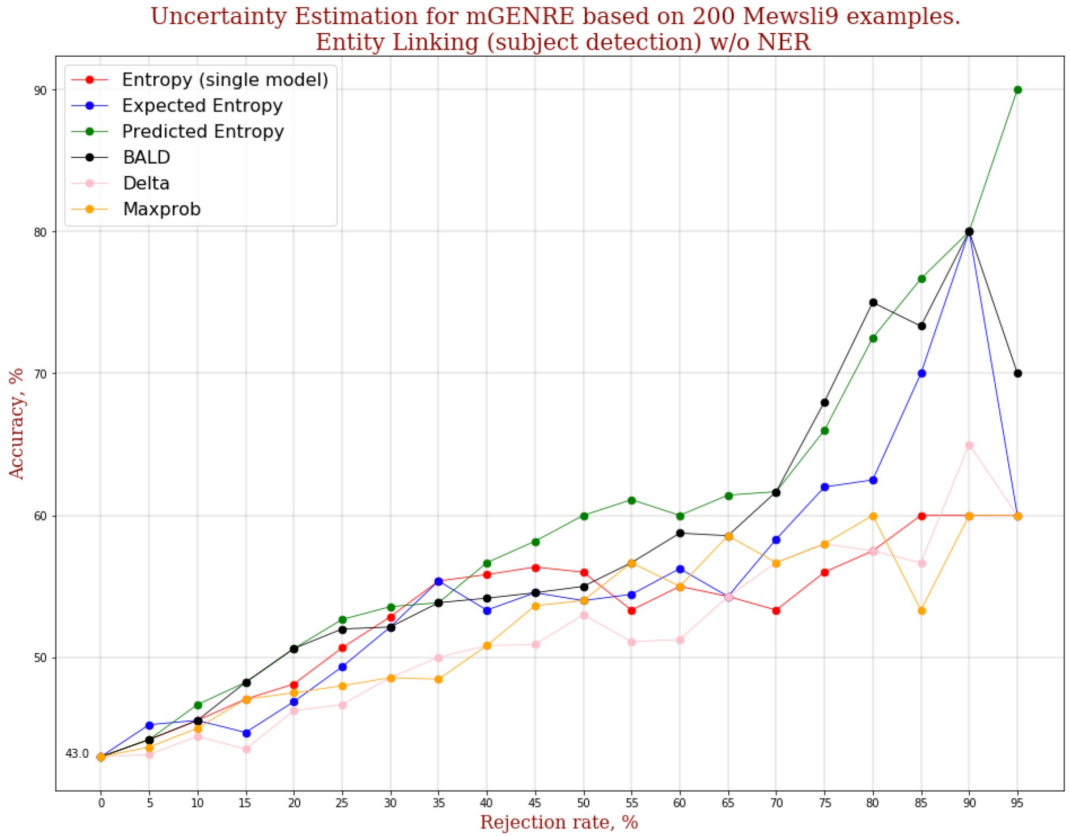




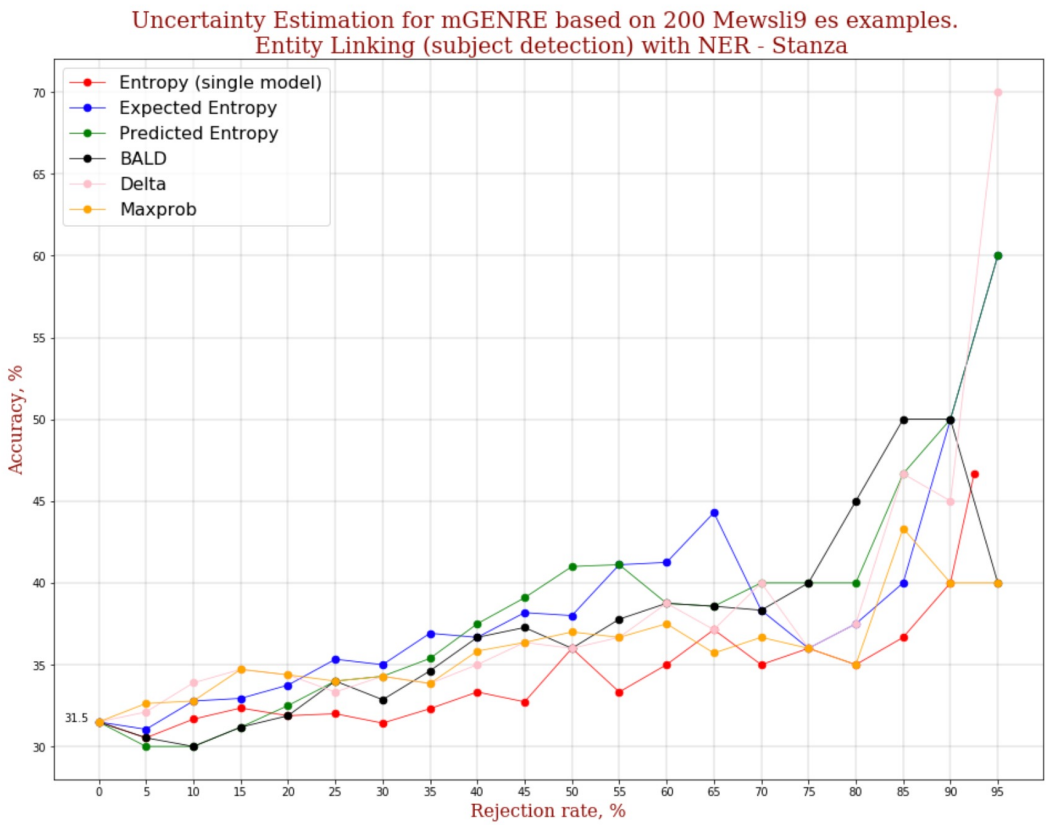
RUBQ 2.0 (English)
Entity Linking. Subject detection

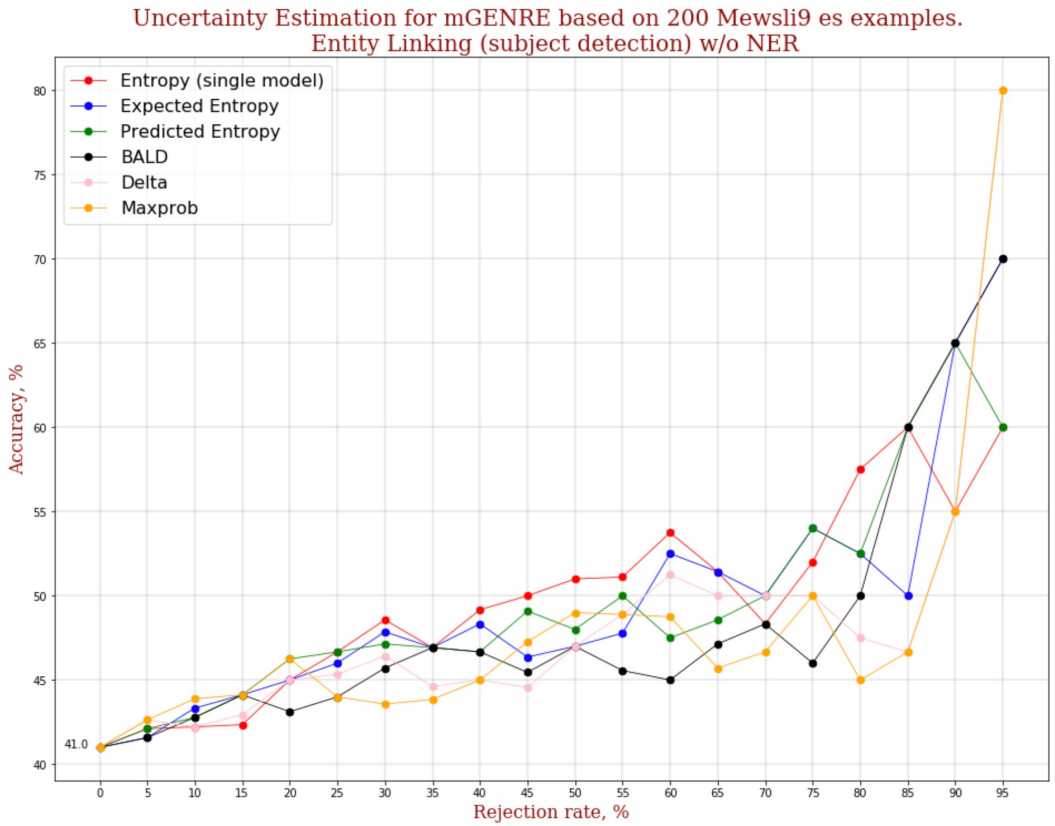
Part 3. Mewsli-9



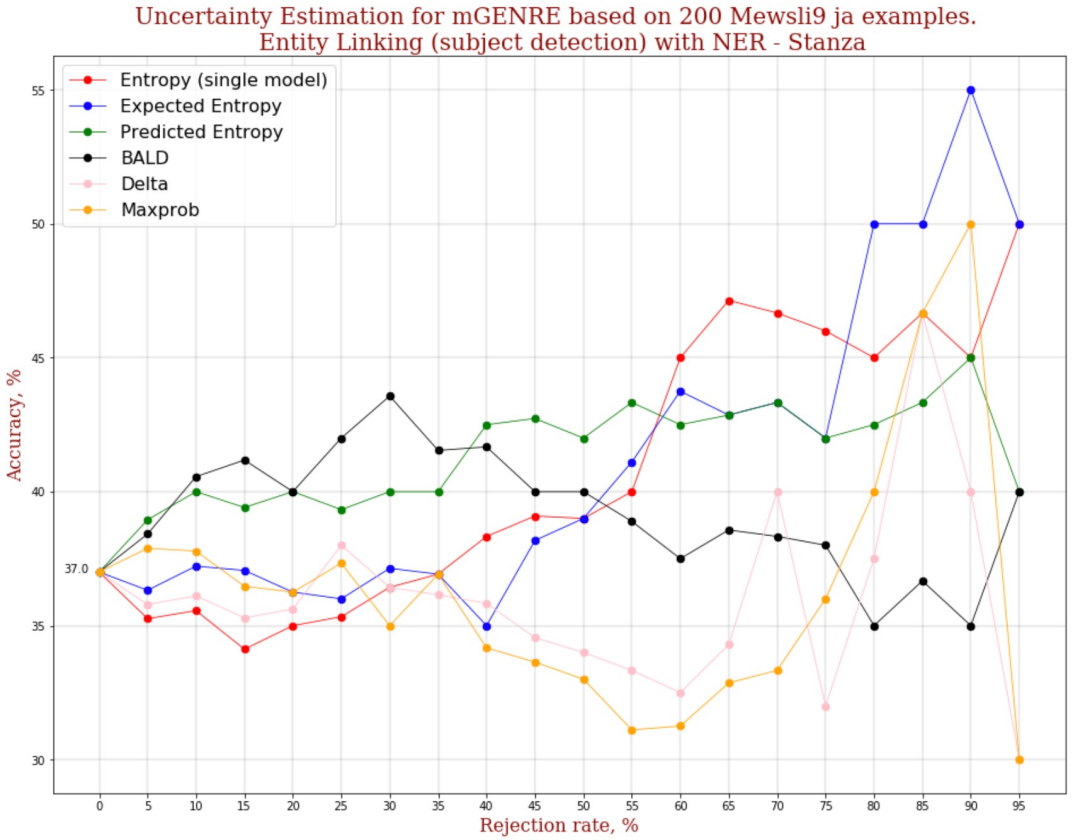


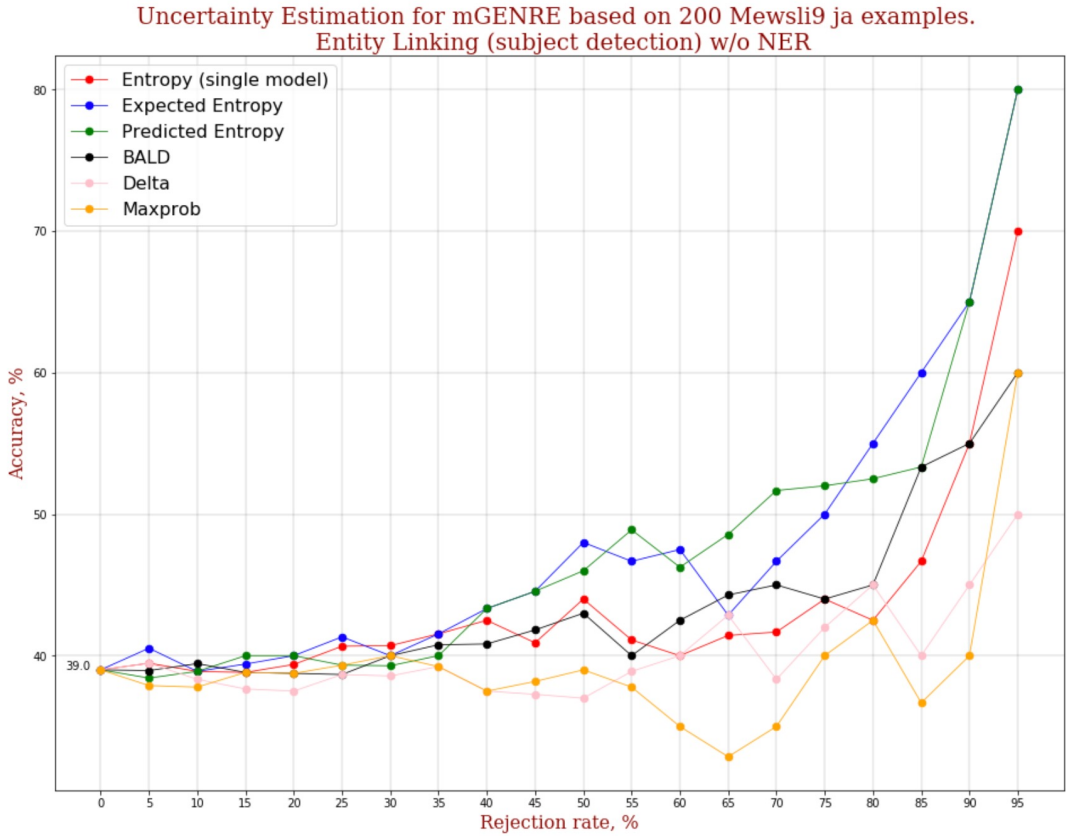
Mewsli-9 (German)
Entity Linking. Subject detection



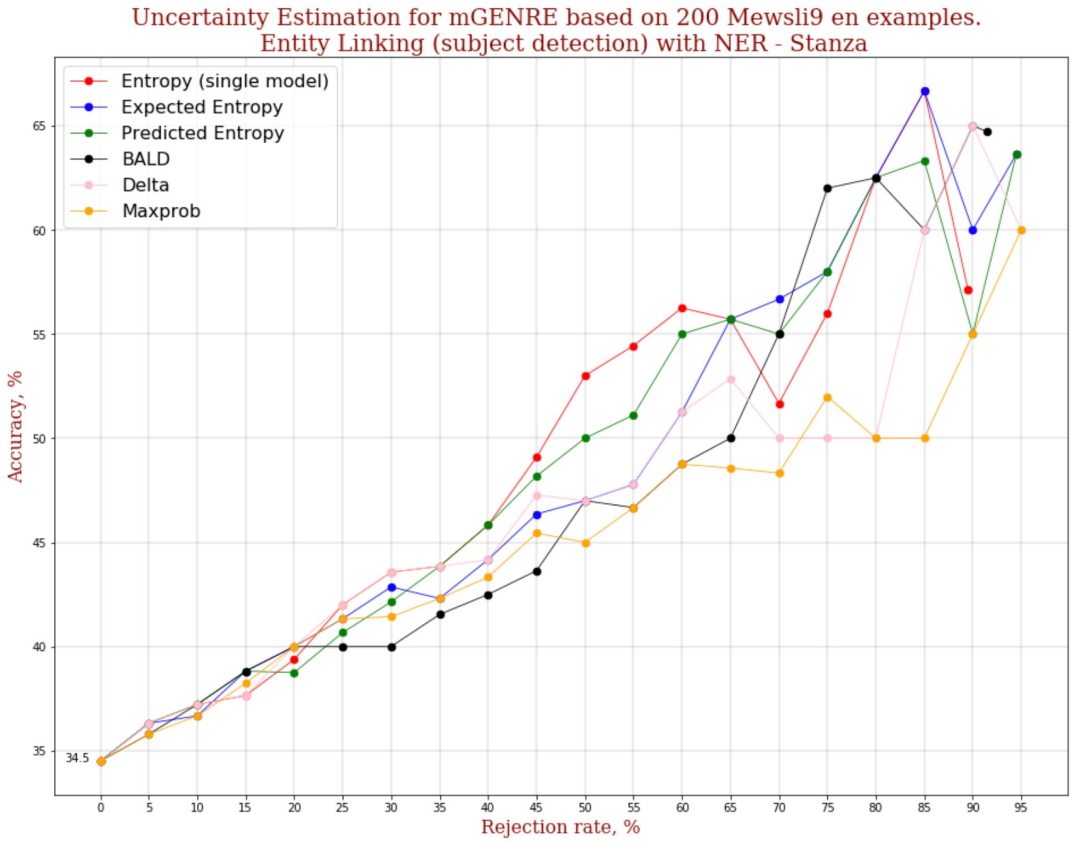


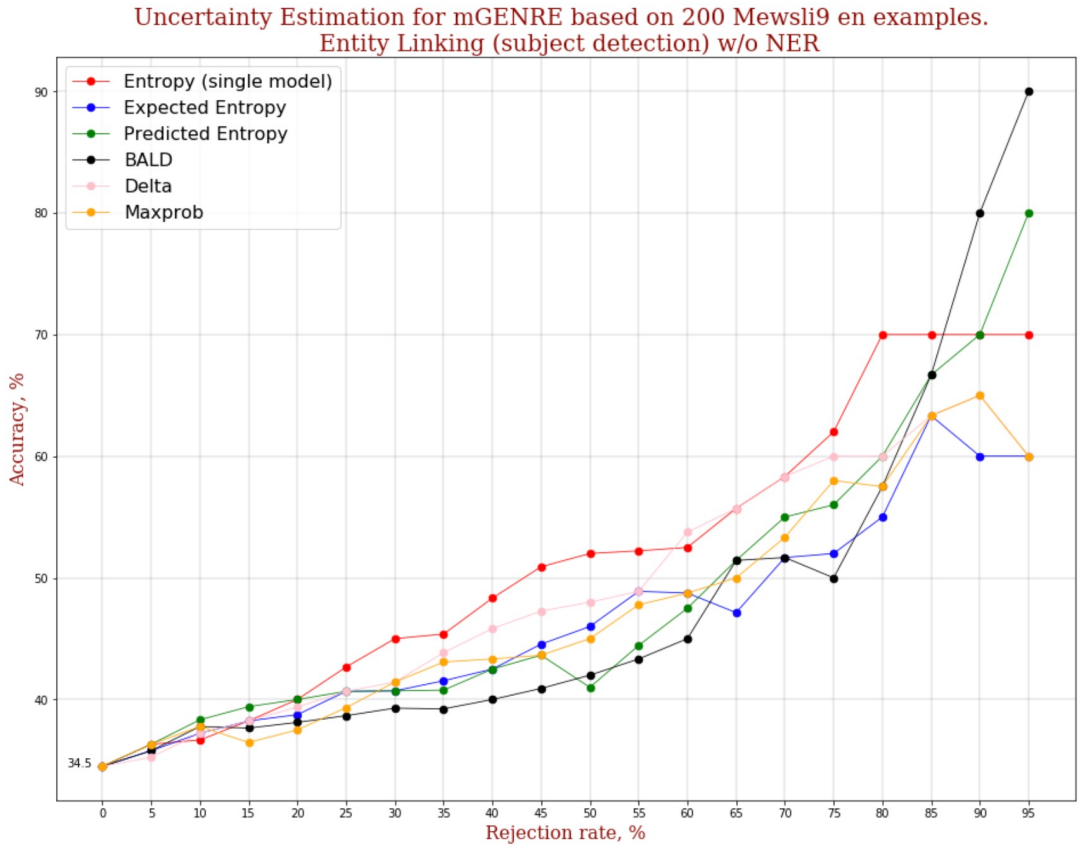
RUBQ 2.0 (Spanish)
Entity Linking. Subject detection



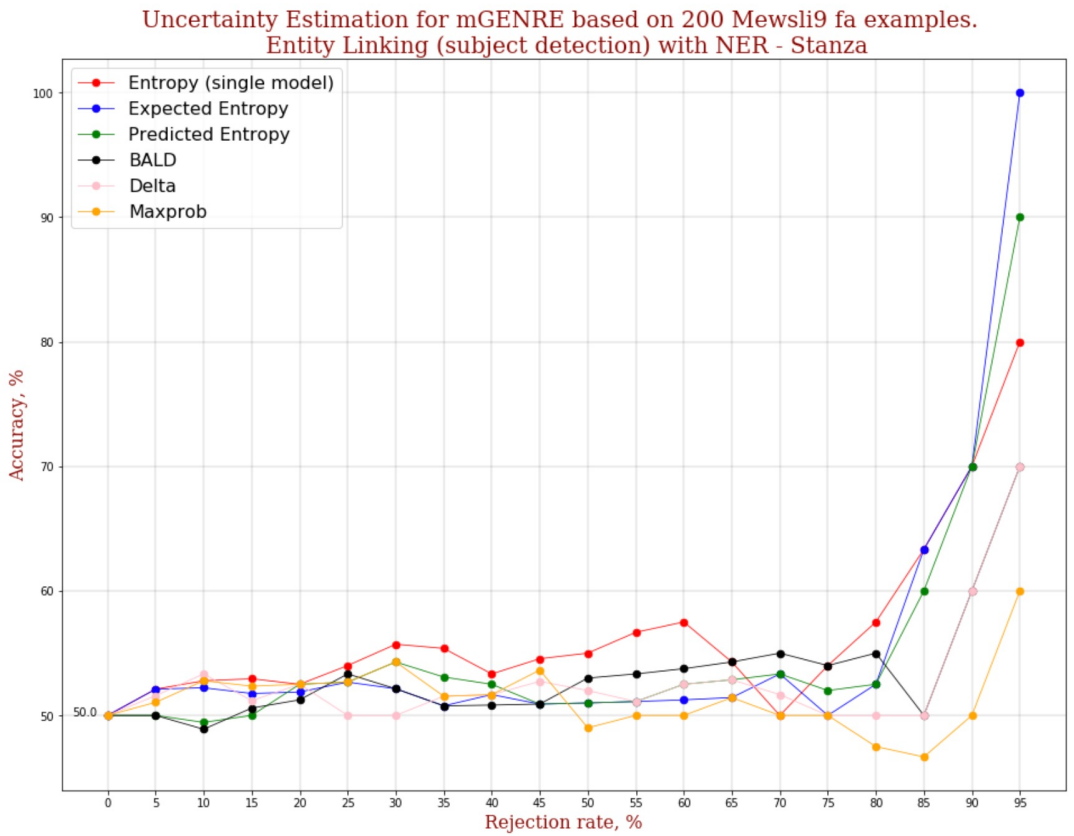


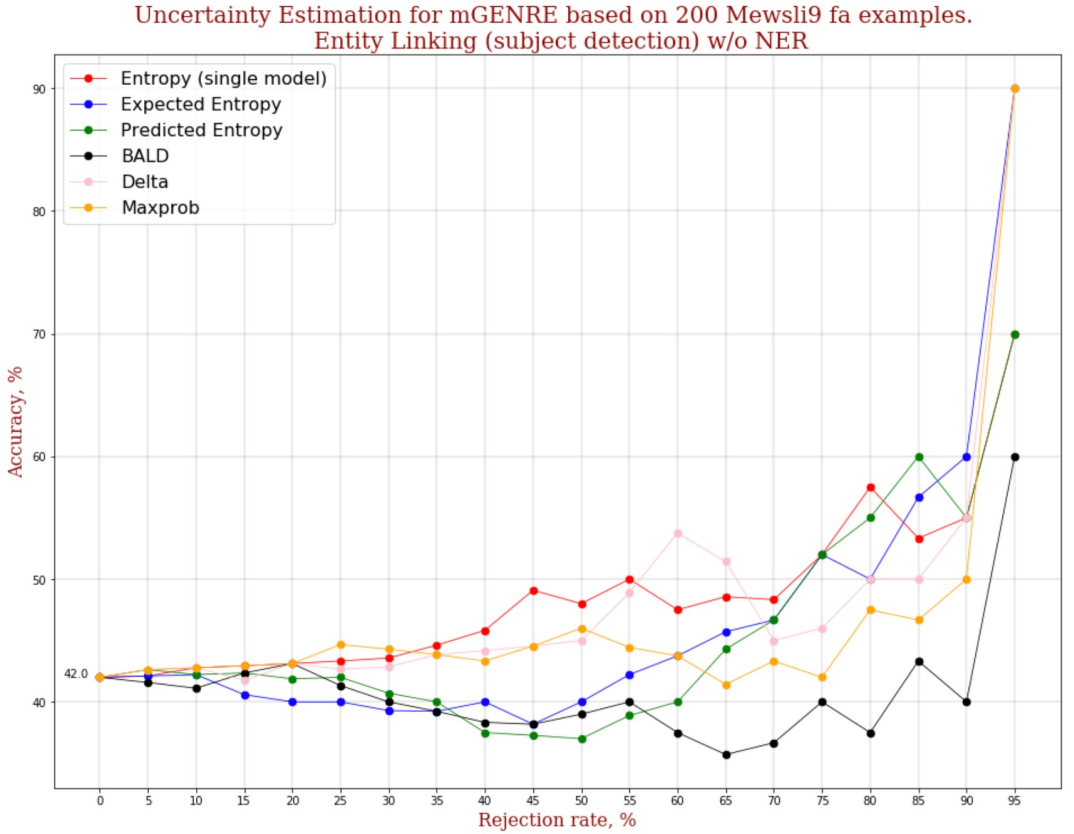
RUBQ 2.0 (Japanese)
Entity Linking. Subject detection



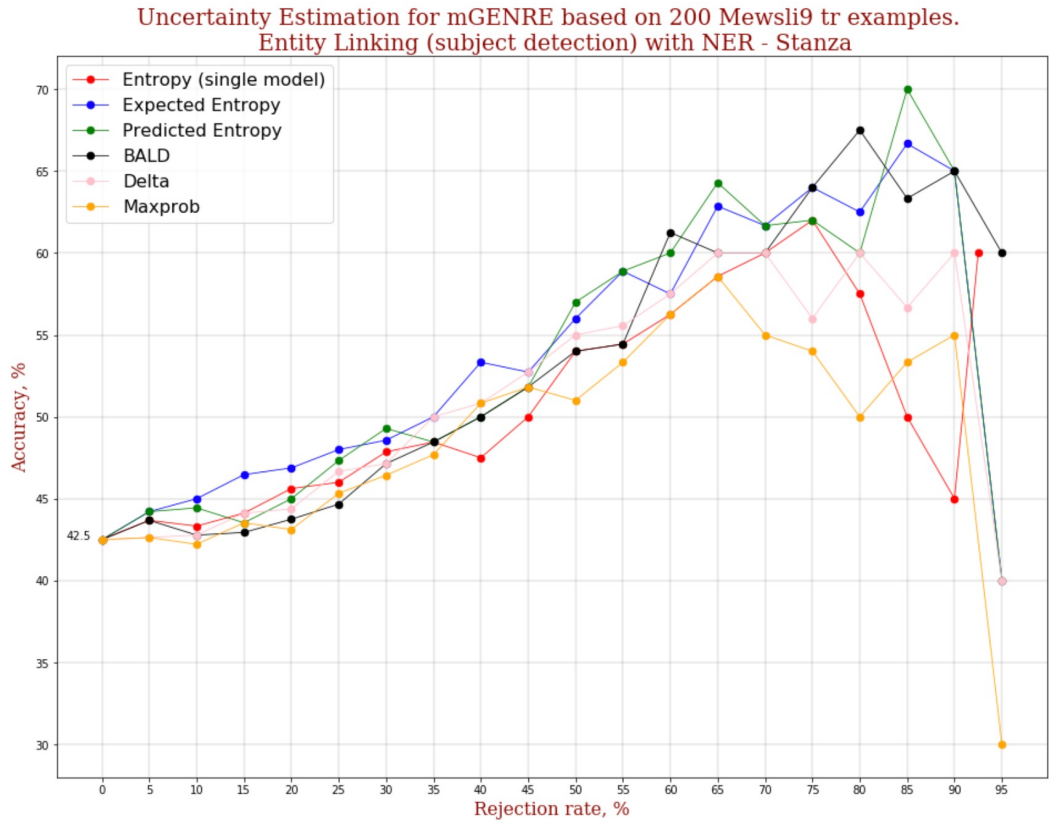


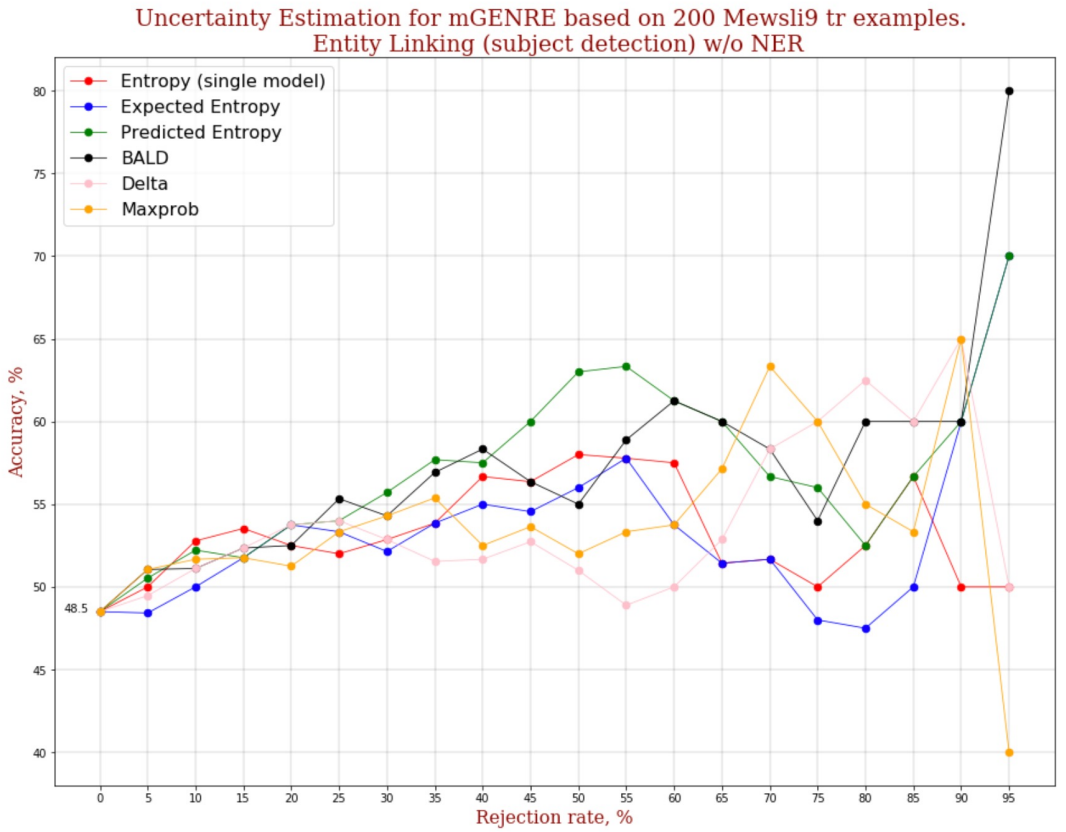
RUBQ 2.0 (English)
Entity Linking. Subject detection



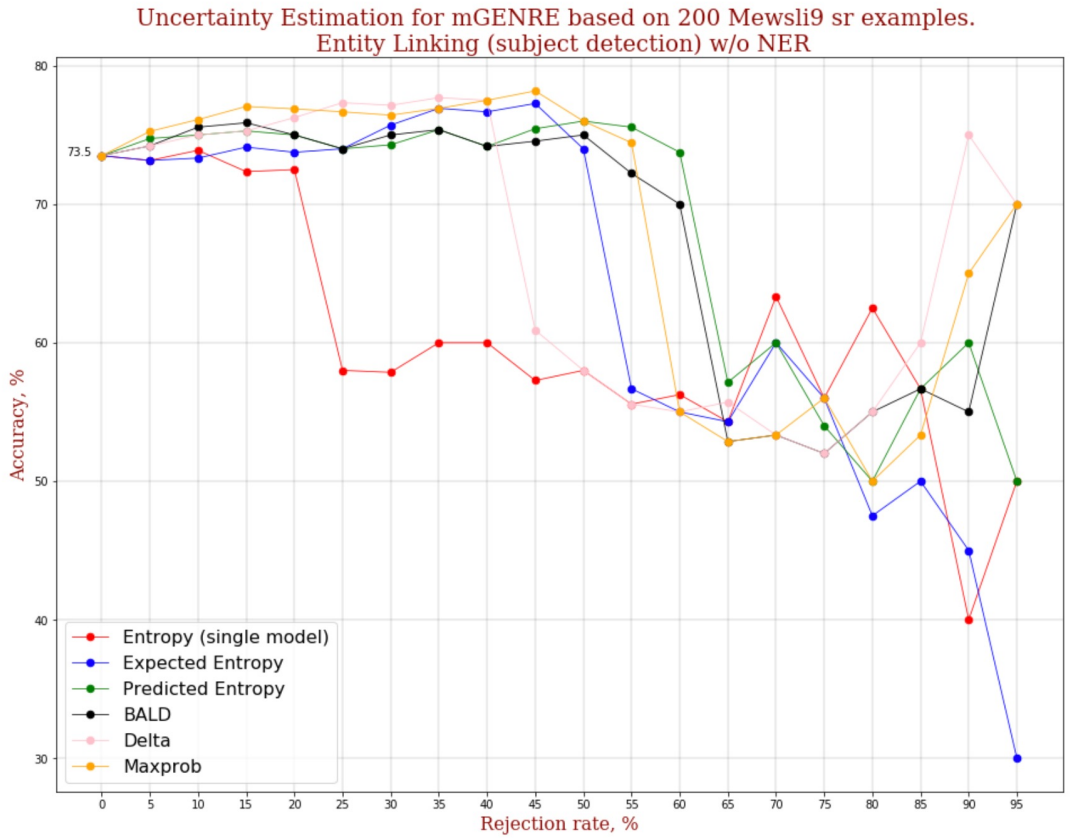


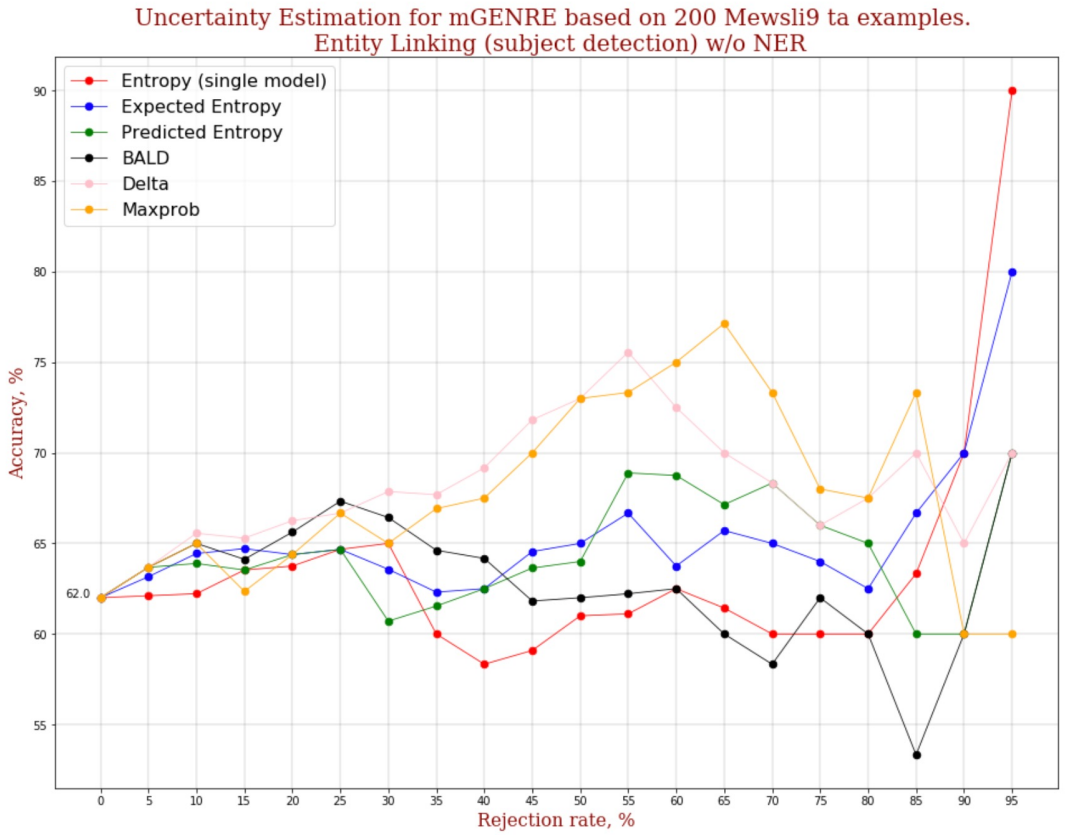
RUBQ 2.0 (Persian)
Entity Linking. Subject detection





RUBQ 2.0 (Turkish)
Entity Linking. Subject detection





RUBQ 2.0 (Serbian & Javanese) Entity Linking. Subject detection

Area under Rejection curve as a measure of uncertainty metric quality (total area = Absolute)

Numbers in table = Area under rejection curve * 100

Dataset	Language	EL/QA	With/without NER	Number of answers	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected Entropy	MEAN
Mewsli 9	German	EL	with_Stanza	full	50.81	53.61	53.17	53.18	52.9	52.71	52.73
			wo_NER	full	44.65	48.36	47.36	42.8	43.99	45.27	45.4
	English	EL	with_Stanza	full	40.75	40.4	38.93	38.41	37.31	39.92	39.29
			wo_NER	full	41.93	38.42	36.95	40.14	38.41	37.92	38.96
	Spanish	EL	with_Stanza	full	28.49	31.12	30.71	30.45	30.04	31.24	30.34
			wo_NER	full	41.43	40.64	38.99	39.36	38.92	40.51	39.98
	Persian	EL	with_Stanza	full	46.25	44.28	44.35	43.73	43.44	44.17	44.37
			wo_NER	full	39.85	36.57	33.71	38.72	37.25	36.57	37.11
	Japanese	EL	with_Stanza	full	33.83	35.08	33.6	30.46	30.24	33.78	32.83
			wo_NER	full	35.02	37.79	35.4	33.39	32.37	37.79	35.29
	Serbian	EL	wo_NER	full	52.8	59.24	58.46	56.13	58.6	57.01	57.04
	Javanese	EL	wo_NER	full	52.37	54.88	53.38	58.14	58.32	54.56	55.28
	Turkish	EL	with_Stanza	full	43.28	45.21	44.47	43.75	41.98	45.66	44.06
			wo_NER	full	45.6	47.92	47.5	45.37	46.02	44.41	46.14
RUBQ 2.0	Russian	EL	wo_NER	single_answer	57.71	60.01	58.42	57.91	56.42	60	58.41
	English	EL	with_Stanza	single_answer	50.26	53.62	51.53	53.24	51.55	52.36	52.09
			wo_NER	single_answer	54.73	57.16	54.67	54.28	54.67	58.31	55.64
	Russian	EL	with_Stanza	single_answer	50.72	55	53.81	53.76	54.38	58.96	54.44
Simple Questions	English	EL	With_Stanza	full	61.54	68.09	67.15	65.14	64.06	11.35	56.22
				single_answer	59.1	63.13	61.37	57.62	56.85	62.38	60.08
			wo_NER	full	22.81	23.04	22.99	21.74	21.56	21.05	22.2
				single_answer	12.79	13.19	11.32	10.28	10.55	14.63	12.13
		QA	With_Stanza	full	4.4	3.67	3.63	3.1	3	3.74	3.59
				single_answer	7.78	7.04	8.45	11.68	11.38	7.18	8.92
			wo_NER	single_answer	9.02	10.69	11.2	10.21	11.32	7.36	9.97
				full	4.67	3.21	2.66	4.53	3.71	4.21	3.83

Language	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected Entropy
English	42.99	44.63	43.11	42.61	41.87	37.24
German	47.73	50.98	50.26	47.99	48.44	48.99
Japanese	34.42	36.44	34.50	31.92	31.30	35.78
Javanese	52.37	54.88	53.38	58.14	58.32	54.56
Persian	43.05	40.42	39.03	41.22	40.34	40.37
Russian	54.22	57.50	56.12	55.83	55.40	59.48
Serbian	52.80	59.24	58.46	56.13	58.60	57.01
Spanish	34.96	35.88	34.85	34.90	34.48	35.88
Turkish	44.44	46.56	45.98	44.56	44.00	45.04

Dataset	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected Entropy
Mewsli 9	42.65	43.82	42.64	42.43	42.13	42.97
RUBQ 2.0	53.36	56.45	54.61	54.80	54.26	57.41
Simple Questions	39.06	41.86	40.71	38.70	38.26	27.35

Absolute AUC Statistics

Area under Rejection curve as a measure of uncertainty metric quality (area added by uncertainty = Comparative)

Dataset	Language	EL/QA	With/without NER	Number of answers	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected entropy	MEAN
Mewsli 9	German	EL	with_Stanza	full	8.73	11.53	11.09	11.1	10.82	10.63	10.65
			wo_NER	full	8.1	11.81	10.81	6.25	7.44	8.72	8.86
	English	EL	with_Stanza	full	11.43	11.08	9.61	9.09	7.99	10.6	9.97
			wo_NER	full	12.61	9.1	7.63	10.82	9.09	8.6	9.64
	Spanish	EL	with_Stanza	full	1.71	4.34	3.93	3.67	3.26	4.46	3.56
			wo_NER	full	6.58	5.79	4.14	4.51	4.07	5.66	5.12
	Persian	EL	with_Stanza	full	3.75	1.78	1.85	1.23	0.94	1.67	1.87
			wo_NER	full	4.15	0.87	-1.99	3.02	1.55	0.87	1.41
	Japanese	EL	with_Stanza	full	2.38	3.63	2.15	-0.99	-1.21	2.33	1.38
			wo_NER	full	1.87	4.64	2.25	0.24	-0.78	4.64	2.14
	Serbian	EL	wo_NER	full	-9.68	-3.24	-4.02	-6.35	-3.88	-5.47	-5.44
	Javanese	EL	wo_NER	full	-0.33	2.18	0.68	5.44	5.62	1.86	2.58
	Turkish	EL	with_Stanza	full	7.16	9.09	8.35	7.63	5.86	9.54	7.94
			wo_NER	full	4.38	6.7	6.28	4.15	4.8	3.19	4.92
RUBQ 2.0	Russian	EL	wo_NER	single_answer	8.84	11.14	9.55	9.04	7.55	11.13	9.54
	English	EL	with_Stanza	single_answer	9.88	13.24	11.15	12.86	11.17	11.98	11.71
			wo_NER	single_answer	12.65	15.08	12.59	12.2	12.59	16.23	13.56
	Russian	EL	with_Stanza	single_answer	9.92	14.2	13.01	12.96	13.58	18.16	13.64
Simple Questions	English	EL	With_Stanza	full	14.36	20.91	19.97	17.96	16.88	20.6	18.45
				single_answer	17.02	21.05	19.29	15.54	14.77	20.3	18
			wo_NER	full	7.51	7.74	7.69	6.44	6.26	5.75	6.9
				single_answer	3.01	3.41	1.54	0.5	0.77	4.85	2.35
		QA	With_Stanza	full	-1.98	-2.71	-2.75	-3.28	-3.38	-2.64	-2.79
				single_answer	-7.52	-8.26	-6.85	-3.62	-3.92	-8.12	-6.38
			wo_NER	single_answer	-4.58	-2.91	-2.4	-3.39	-2.28	-6.24	-3.63
				full	-0.43	-1.89	-2.44	-0.57	-1.39	-0.89	-1.27

Comparative AUC Statistics

Language	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected Entropy
English	11.06	12.70	11.18	10.68	9.94	12.36
German	8.41	11.67	10.95	8.68	9.13	9.68
Japanese	2.12	4.14	2.20	-0.38	-1.00	3.49
Javanese	-0.33	2.18	0.68	5.44	5.62	1.86
Persian	3.95	1.32	-0.07	2.12	1.25	1.27
Russian	9.38	12.67	11.28	11.00	10.56	14.65
Serbian	-9.68	-3.24	-4.02	-6.35	-3.88	-5.47
Spanish	4.14	5.06	4.04	4.09	3.66	5.06
Turkish	5.77	7.90	7.32	5.89	5.33	6.36

Dataset	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected Entropy
Mewsli 9	4.49	5.66	4.48	4.27	3.97	4.81
RUBQ 2.0	10.32	13.42	11.58	11.76	11.22	14.38
Simple Questions	10.48	13.28	12.12	10.11	9.67	12.88