Joint Posterior Revision of NLP Annotations via Ontological Knowledge [ADDENDUM]

Marco Rospocher **Francesco Corcoglioniti**
Fondazione Bruno Kessler (FBK-irst)
{rospocher, corcoglio}@fbk.eu

COMPLEMENTARY TABLES WITH THE SCORES FOR ALL METRICS AND EVALUATION MEASURES CONSIDERED IN THE PAPER:

Marco Rospocher, Francesco Corcoglioniti.
Joint Posterior Revision of NLP Annotations via Ontological Knowledge.
In Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence, IJCAI-ECAI 2018, Stockholm, Sweden, July 13-19, 2018

Table 1: Precision, recall, and F_1 scores (**micro-averaged, only gold mentions**) for type, link, and type+link measures for both considered settings on the three evaluation datasets. Score differences between the settings (*with* JPARK vs. *standard*) are reported.

		type			link			type+link		
		P	R	F_1	P	R	F_1	P	R	F_1
AIDA	*standard*	94.30%	87.50%	90.80%	66.20%	65.20%	65.60%	63.40%	62.50%	63.00%
	with JPARK	95.00%	88.10%	91.40%	67.10%	65.40%	66.20%	65.50%	63.70%	64.60%
	Δ	0.70%	0.60%	0.60%	0.90%	0.20%	0.60%	2.10%	1.20%	1.60%
MEANTIME	*standard*	88.20%	69.50%	77.70%	70.30%	55.60%	62.10%	63.50%	50.20%	56.10%
	with JPARK	91.40%	72.00%	80.50%	70.50%	55.70%	62.20%	67.00%	53.00%	59.20%
	Δ	3.20%	2.50%	2.80%	0.20%	0.10%	0.10%	3.50%	2.80%	3.10%
TAC-KBP	*standard*	91.10%	65.20%	76.00%	40.10%	42.30%	41.20%	36.70%	38.60%	37.60%
	with JPARK	92.60%	66.30%	77.20%	41.20%	42.60%	41.90%	38.90%	40.20%	39.50%
	Δ	1.50%	1.10%	1.20%	1.10%	0.30%	0.70%	2.20%	1.60%	1.90%

Table 2: Precision, recall, and F_1 scores (**micro-averaged, considering all systems' mentions**) for type, link, and type+link measures for both considered settings on the three evaluation datasets. Score differences between the settings (*with* JPARK vs. *standard*) are reported. *Note: Precision and F_1 scores on TAC-KBP are extremely low (and ignored in the summary reported in the paper) due to the fact that only mentions of one single entity per document is annotated in the gold standard (4969 mentions), while the tools annotate all the content of each document (116009 mentions).*

		type			link			type+link		
		P	R	F_1	P	R	F_1	P	R	F_1
AIDA	*standard*	88.10%	87.50%	87.80%	63.40%	65.20%	64.20%	60.80%	62.50%	61.60%
	with JPARK	88.70%	88.10%	88.40%	64.40%	65.40%	64.90%	62.70%	63.70%	63.20%
	Δ	0.60%	0.60%	0.60%	1.00%	0.20%	0.70%	1.90%	1.20%	1.60%
MEANTIME	*standard*	48.40%	69.50%	57.00%	43.00%	55.60%	48.50%	38.80%	50.20%	43.80%
	with JPARK	50.10%	72.00%	59.10%	43.20%	55.70%	48.60%	41.00%	53.00%	46.20%
	Δ	1.70%	2.50%	2.10%	0.20%	0.10%	0.10%	2.20%	2.80%	2.40%
TAC-KBP	*standard*	*2.80%*	65.20%	*5.40%*	*1.20%*	42.30%	*2.30%*	*1.10%*	38.60%	*2.10%*
	with JPARK	*2.80%*	66.30%	*5.40%*	*1.20%*	42.60%	*2.40%*	*1.10%*	40.20%	*2.20%*
	Δ	*0.00%*	1.10%	*0.00%*	*0.00%*	0.30%	*0.10%*	*0.00%*	1.60%	*0.10%*

Table 3: Precision, recall, and F_1 scores (**macro-averaged by document, only gold mentions**) for type, link, and type+link measures for both considered settings on the three evaluation datasets. Score differences between the settings (*with* JPARK vs. *standard*) are reported.

		type			link			type+link		
		P	R	F_1	P	R	F_1	P	R	F_1
AIDA	*standard*	93.90%	86.10%	89.50%	64.60%	65.10%	63.60%	61.60%	62.20%	60.70%
	with JPARK	94.70%	86.90%	90.20%	65.60%	65.40%	64.30%	63.60%	63.50%	62.40%
	Δ	0.80%	0.80%	0.70%	1.00%	0.30%	0.70%	2.00%	1.30%	1.70%
MEANTIME	*standard*	88.60%	71.80%	77.60%	68.40%	55.10%	59.50%	63.00%	50.90%	54.80%
	with JPARK	91.40%	74.10%	80.10%	68.50%	55.10%	59.50%	66.10%	53.40%	57.60%
	Δ	2.80%	2.30%	2.50%	0.10%	0.00%	0.00%	3.10%	2.50%	2.80%
TAC-KBP	*standard*	81.10%	69.50%	73.30%	26.00%	23.10%	24.10%	23.30%	20.70%	21.60%
	with JPARK	83.00%	71.10%	75.00%	26.20%	23.30%	24.20%	24.60%	21.90%	22.80%
	Δ	1.90%	1.60%	1.70%	0.20%	0.20%	0.10%	1.30%	1.20%	1.20%

Table 4: Precision, recall, and F_1 scores (**macro-averaged by NERC type, only gold mentions**) for type, link, and type+link measures for both considered settings on the three evaluation datasets. Score differences between the settings (*with* JPARK vs. *standard*) are reported.

		type			link			type+link		
		P	R	F_1	P	R	F_1	P	R	F_1
AIDA	*standard*	93.70%	86.30%	89.80%	60.10%	58.40%	59.20%	60.10%	58.40%	59.20%
	with JPARK	94.50%	87.00%	90.50%	62.20%	59.70%	60.90%	62.20%	59.70%	60.90%
	Δ	0.80%	0.70%	0.70%	2.10%	1.30%	1.70%	2.10%	1.30%	1.70%
MEANTIME	*standard*	96.00%	68.00%	79.50%	60.60%	43.70%	50.80%	60.60%	43.70%	50.80%
	with JPARK	97.70%	69.30%	81.00%	62.20%	45.30%	52.30%	62.20%	45.30%	52.30%
	Δ	1.70%	1.30%	1.50%	1.60%	1.60%	1.50%	1.60%	1.60%	1.50%
TAC-KBP	*standard*	92.40%	66.60%	77.20%	36.80%	39.70%	38.10%	36.80%	39.70%	38.10%
	with JPARK	94.00%	68.10%	78.50%	38.60%	41.20%	39.70%	38.60%	41.20%	39.70%
	Δ	1.60%	1.50%	1.30%	1.80%	1.50%	1.60%	1.80%	1.50%	1.60%