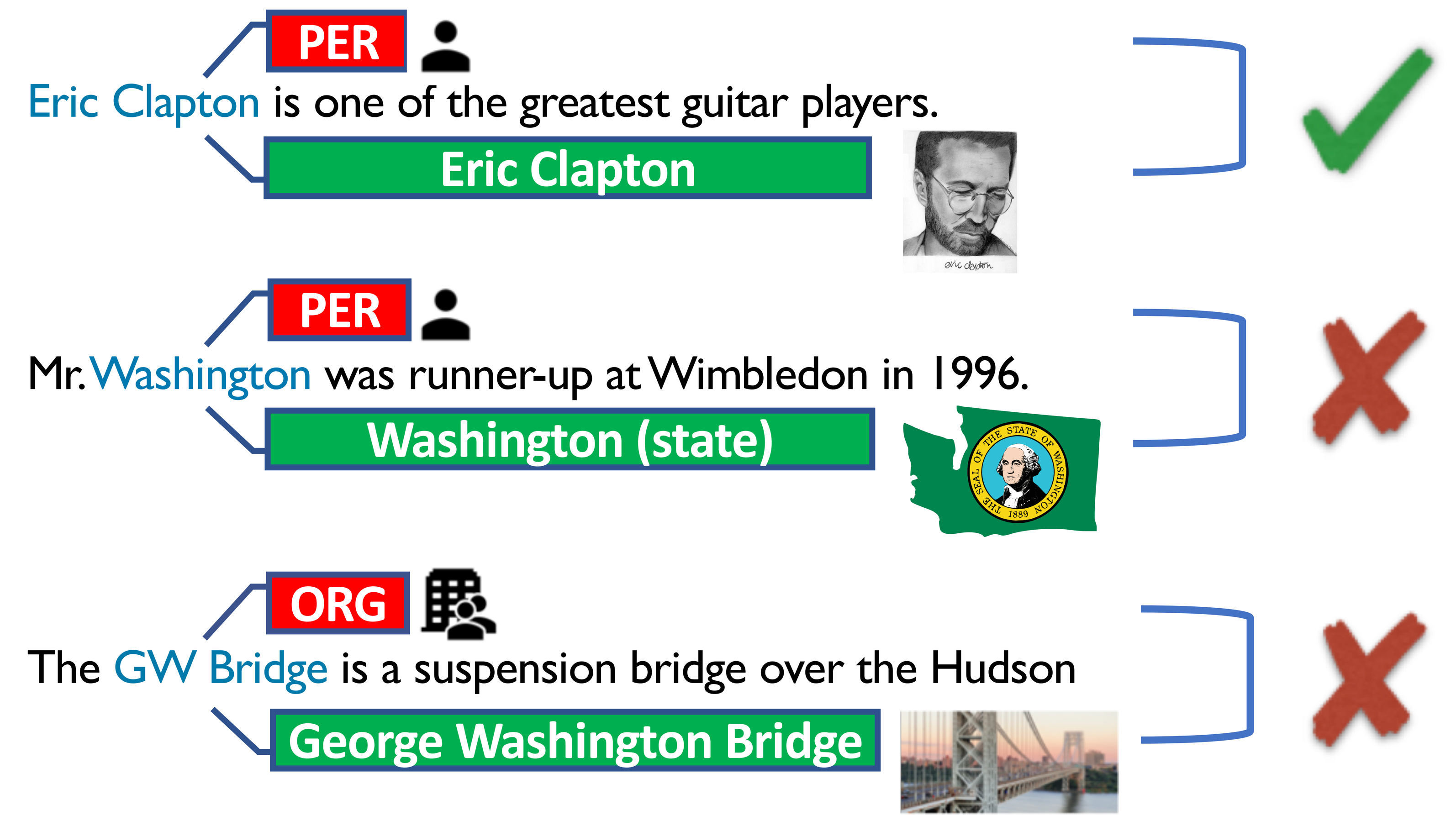
Joint Posterior Revision of NLP Annotations via Ontological Knowledge

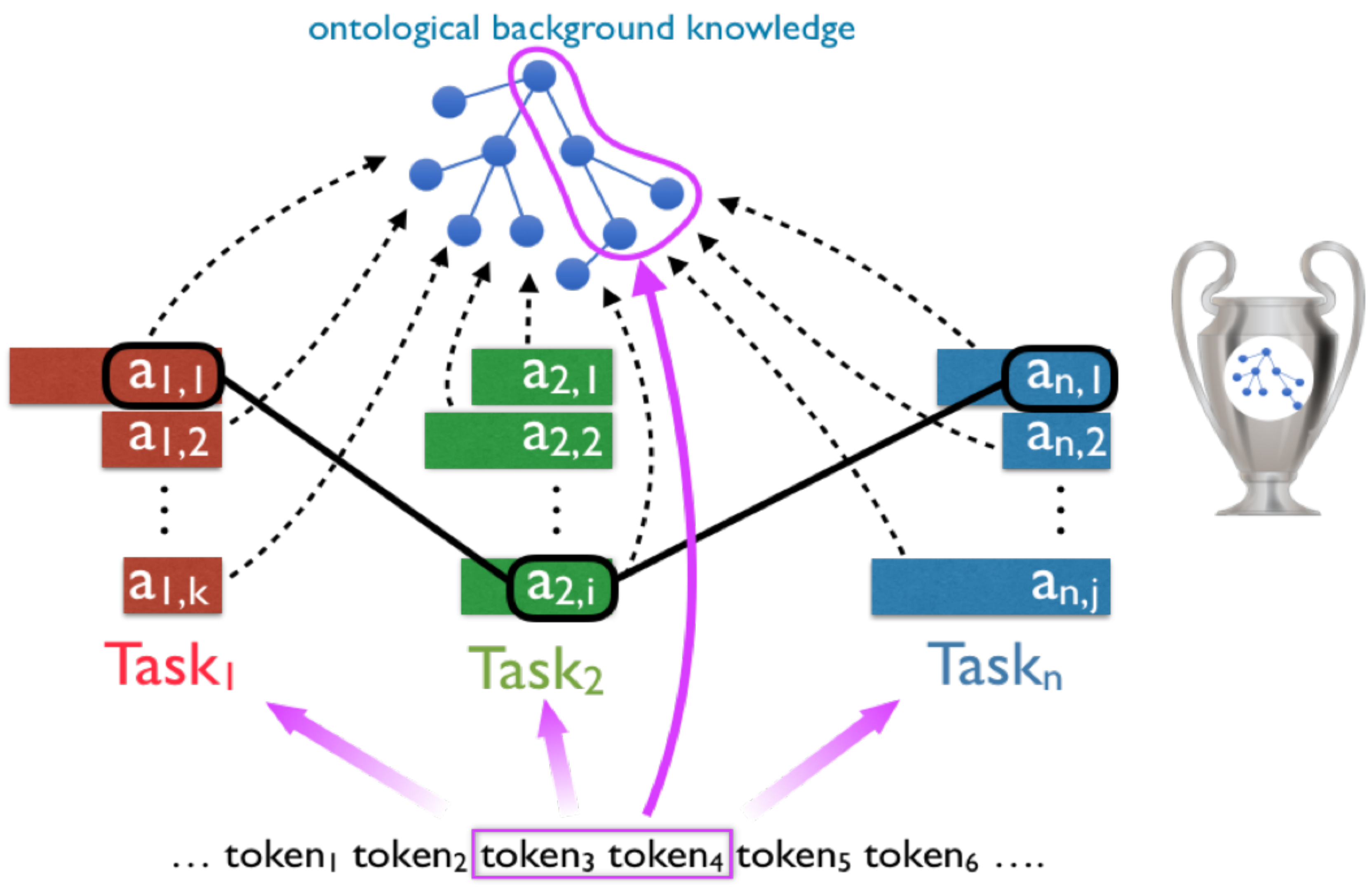
Marco Rospocher (rospocher@fbk.eu), Francesco Corcoglioniti (corcoglio@fbk.eu)

JPARK

1. Problem: Incoherent Mention-level NLP Annotations



2. Solution: Coherence via Ontology



3. General Probabilistic Model

Variables

- m entity mention
- $\boldsymbol{a} = (a_1 \dots a_n)$ NLP annotations
- B NLP Background Knowledge
- K "The" Ontological Knowledge
- C Entity's ontological class set (from K)

Conditional Independence Assumptions

1. $\mathrm{P}(\boldsymbol{a}|m,B,K,C) = \prod_i \mathrm{P}(a_i|m,B,K,C)$
2. $\mathrm{P}(a_i|m,B,K) = \mathrm{P}(a_i|m,B)$
3. $\mathrm{P}(C|a_i,m,B,K) = \mathrm{P}(C|a_i,K)$

$$= \arg\max_{\boldsymbol{a}} \mathrm{P}(\boldsymbol{a}|m,B,K)$$

$$\mathrm{P}(\boldsymbol{a}|m,B,K) \overset{\mathrm{M}}{=} \sum_C \mathrm{P}(\boldsymbol{a},C|m,B,K)$$

$$\mathrm{P}(\boldsymbol{a},C|m,B,K) \overset{\mathrm{CP}}{=} \mathrm{P}(C|m,B,K)\cdot\mathrm{P}(\boldsymbol{a}|m,B,K,C)$$

$$\overset{①}{=} \mathrm{P}(C|m,B,K)\cdot\prod_i \mathrm{P}(a_i|m,B,K,C)$$

$$\overset{\mathrm{CP}}{=} \frac{\prod_i \mathrm{P}(a_i,C|m,B,K)}{\mathrm{P}(C|m,B,K)^{n-1}}$$

$$\mathrm{P}(C|m,B,K) \overset{\mathrm{M}^*}{=} \Big(\prod_i \sum_{a_i} \mathrm{P}(a_i,C|m,B,K)\Big)^{\frac{1}{n}}$$

$$\mathrm{P}(a_i,C|m,B,K) \overset{\mathrm{CP}}{=} \mathrm{P}(a_i|m,B,K)\cdot\mathrm{P}(C|a_i,m,B,K)$$

$$\overset{②③}{=} \underbrace{\mathrm{P}(a_i|m,B)}_{\text{confidence score}}\cdot\underbrace{\mathrm{P}(C|a_i,K)}_{\text{learned from data}}$$

4. Model Instantiated on NERC + EL

Ontological Background Knowledge

YAGO select knowledge: 6,016,695 entities; Taxonomy of 568,255 classes + Wikipedia The Free Encyclopedia (only ingoing links)

Estimating NERC

$$\mathrm{P}(C|a_{\mathrm{NERC}},K) = \alpha\cdot\mathrm{P}(C|K) + (1-\alpha)\cdot\mathrm{P}(C|a_{\mathrm{NERC}},G)$$

$$\mathrm{P}(C|K) = \frac{n_K(C)}{\sum_{C'} n_K(C')}$$

Prior (popularity based on entity ingoing links)

$$\mathrm{P}(C|a_{\mathrm{NERC}},G) = \frac{n_G(C,a_{\mathrm{NERC}})}{\sum_{C'} n_G(C',a_{\mathrm{NERC}})}$$

Computed from gold corpus (NERC + class annotations)

Consider only class sets restricted to popular classes (seen at least n^* times in the gold corpus)

Estimating EL

$$\mathrm{P}(C|a_{\mathrm{EL}},K) = \mathbf{1}_{\{C_K(a_{\mathrm{EL}})\}}(C)$$

Deterministically computable from the classes of the entity (possibly leveraging alignments between EL Knowledge Base and YAGO)

5. Evaluation

Tools: **Stanford CoreNLP** DBpedia Spotlight

Gold Corpus (NERC): AIDA CoNLL-YAGO (train)

Datasets: ① AIDA CoNLL-YAGO (test-b) ② MEANTIME ③ TAC-KBP

Research question: Does the JPARK posterior joint revision of the annotations from Stanford CoreNLP (NERC) and DBpedia Spotlight (EL), via YAGO, improve their performances?

Measures: NERC / EL / NERC+EL

	NERC			EL			NERC+EL		
	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%**

Restricted to gold mentions. Similar improvements also considering all mentions, and macro-averaging by document or by NERC type.

IJCAI ECAI 2018

#4479

FONDAZIONE BRUNO KESSLER