Ryan Smith is a versatile, classically trained pianist from South Carolina. Ryan has performed solo and collaborative recitals across the Southeast. He has recently appeared at the Columbia Museum of Art, Piccolo Spoleto Festival, the Palmetto Theatre, ... [Edit: other venues he played at.]

Ryan is a member of the Columbia Music Teachers Association. In 2007 Ryan was designated a Nationally Certified Teacher of Music (NCTM) by the National Music Teachers' Association. [Edit: continue to discuss his students.]

In addition to performing, ... [Edit: his other interests.]

The Proposed System: Posterior Probability Pursuit

**Sparse Signal**

Only a few **good** features

- “Pianist”, “South Carolina”, “Columbia Music Teachers Association”

Many less/IR-relevant features

- “a member of”, “in addition to”, “versatile”, “southeast”

Key idea: Use only most discriminative features

Sparse signal recovery problem.

Distance metric: Posterior Probability.

Greedily select best features.

**Model Unknown**

- No knowledge base is complete
- Key challenge is to discriminate between entities in our KB and unknown entities

Define: **Out-of-Knowledge-Base entity** $e_u$

$$P(f|e_u) = \frac{|\{e : f \in F_e\}|}{|F_e|}$$

- Treat the “Unknown” entity as a regular entity in the Bayesian formulation.

**Phrases as Features**

Phrases are better features than tokens

- “Columbia Music Teachers Association”, or “classical pianist”

How to obtain phrases?

1. Structured fields in KB
2. Algorithmically detect phrases using phrase language model, captures phrase boundaries as using partially observable markov model.

Evaluation: People Entity Linking at the Tail

**Setup**

- KB with ~100M people entities
- 50 names, retrieve top 20 documents
- Manually label w/exhaustive search
- 555 docs (w/185 matching to KB)
- Methods in comparison:
  - TF-IDF with cosine-similarity
  - Lasso-based method
  - Naive Bayes
  - PPP

Achieves 95% Precision, with over 60% recall