Entity Linking at the Tail: Sparse Signals, Unknown Entities, and Phrase Models

Yuzhe Jin, Emre Kıcım, Kuansan Wang, Ricky Loynd
Microsoft Research

emrek@microsoft.com
Most entities mentioned on the web not famous, not popular

Ryan Smith is a versatile, classically trained pianist from South Carolina. Ryan has performed ...

More confounds, more uniform prior, often fewer features
Briefly...

<table>
<thead>
<tr>
<th>Solution Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse signal processing</td>
</tr>
<tr>
<td>Explicit model of entities outside of KB</td>
</tr>
<tr>
<td>Phrase LM to efficiently capture higher-order statistics</td>
</tr>
</tbody>
</table>

Evaluation: > 95% precision and > 60% recall
1. Sparse Signal Recovery

Only a few good features
• “Pianist”, “South Carolina”, “Columbia Music Teachers Association”

Sparse signal recovery problem
• Bayesian sequential selection
• Distance metric: Posterior Probability

\[ f_k = \arg \min_{g \in G \setminus F_{k-1}} \left[ \min_{e \in E_c} -\log P(e | F_{k-1} \cup \{g\}) \right] \]

• Similar to matching pursuit, but different distance metric
2. Modeling unknown entities

- No knowledge base is complete
- **Key challenge:** discriminate unknown vs known entities

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

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

- Integrate $e_u$ into entity collection for feature selection
3. Phrase language model

Phrases are better features than tokens
• “Columbia Music Teachers Association”, or “classical pianist”

Obtain phrases 2 ways
1. Directly from structured fields in KB
2. Indirectly from free text using phrase LM

Phrase Unigram Language Model
• Treat every phrase as a unigram token
• Detect phrase boundaries using Partially Observable Markov (POM) model
• Captures high-order dependencies among words
• Manageable space compared with traditional high-order LM
Evaluation Setup

• KB with ~100M people entities
• 50 names, retrieve top 20 documents
• Manually label w/exhaustive search
• 555 docs, with 185 matching to KB

• Evaluated combinations of PPP, unknown model, phrase features
• Baselines:
  • TF-IDF with cosine similarity
  • Lasso method
  • Naïve Bayes (with unknown model)
Achieves 96% Precision, with 62% recall
Summary

• We address new challenges of entity linking at the tail:
  • Posterior Probability Pursuit (sparse signal recovery)
  • Model unknown entities
  • Phrase features

Result: 96% precision, 62% recall linking web pages to KB of 100M+ people entities

More details in paper
Contact: Emre Kıcıman, emrek@microsoft.com