Low-Rank Subspaces for Unsupervised Entity Linking

Akhil Arora (akhil.arora@epfl.ch), Alberto García Durán, and Robert West
Data Science Laboratory (DLAB), EPFL

What is Entity Linking?

Michael Jordan is one of the leading figures in machine learning. In 2016, **Science** reported him as the world's most influential computer scientist.

C_{11} = Michael_Jordan_(basketball_player)
C_{12} = Michael_Jordan_(computer_scientist)
C_{13} = Mike_Jordan_(racing_driver)
C_{14} = Michael-Hakim_Jordan

C_{21} = Natural_Science
C_{22} = Applied_Science
C_{23} = Science_(album)
C_{24} = Life_Science
C_{25} = Science_(journal)

Fundamental NLP task with many applications

- Information extraction
- Automatic KB construction
- Enabling network navigation

How is Entity Linking Performed?

- Dictionaries/alias-tables for high-quality candidate generation

Candidate Entity	Prior P(e|m)
Michael_Jordan	0.997521
Michael_I._Jordan	0.000826
Michael_Jordan_statue	0.000826
Michael_Jordan_(footballer)	0.000826

- Supervised learning via informative features
  - ➢ Prior
  - ➢ Local/Global context
- Sophisticated models on labelled data
  - ➢ XGBoost
  - ➢ Deep neural networks

Unaddressed Research Questions

- Absence of annotated/labelled training data
- Ability to operate at Web Scale

Unsupervised Entity Linking

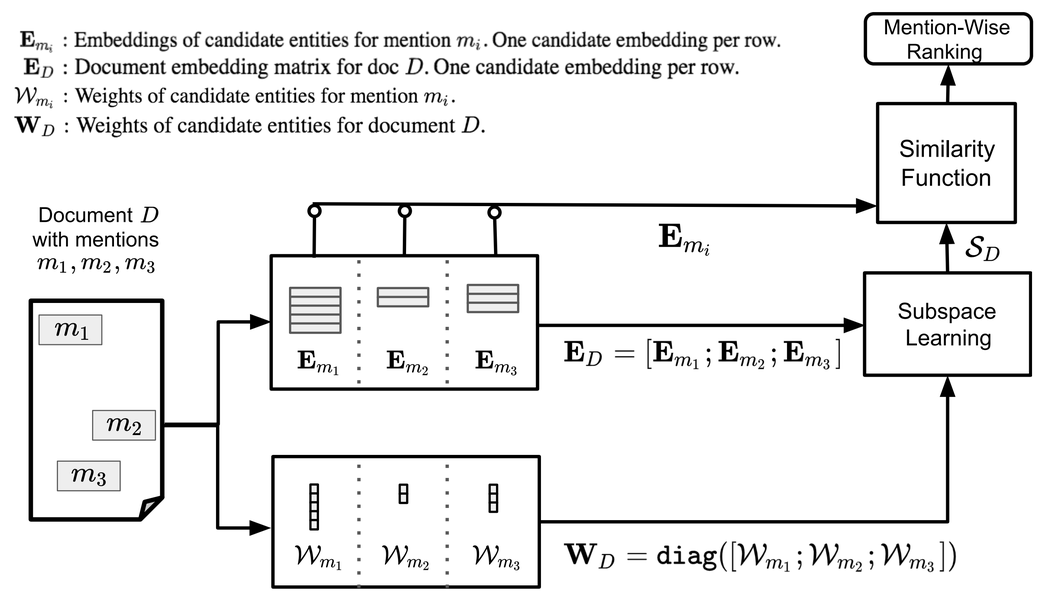
- Absolute "absence" of annotated data
  - Specialized domains: medicine, law, etc.
  - Proprietary KGs
- Accessible resources:
  - List of entity names, or "aliases"
  - Reference KB

Challenges

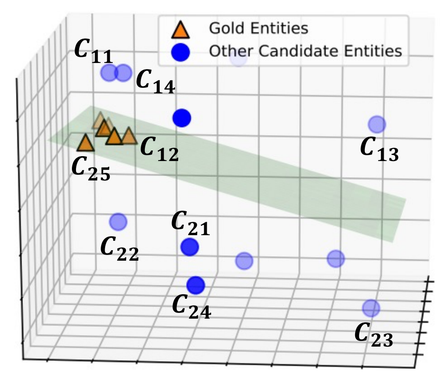
- No module can make use of annotated data!
  - ❌ Candidate generator using dictionaries
  - ❌ Features (e.g. prior probability)
  - ❌ Aligned entity and mention embeddings
  - ❌ Training supervised models

Unsupervised Entity Linking with Eigenthemes

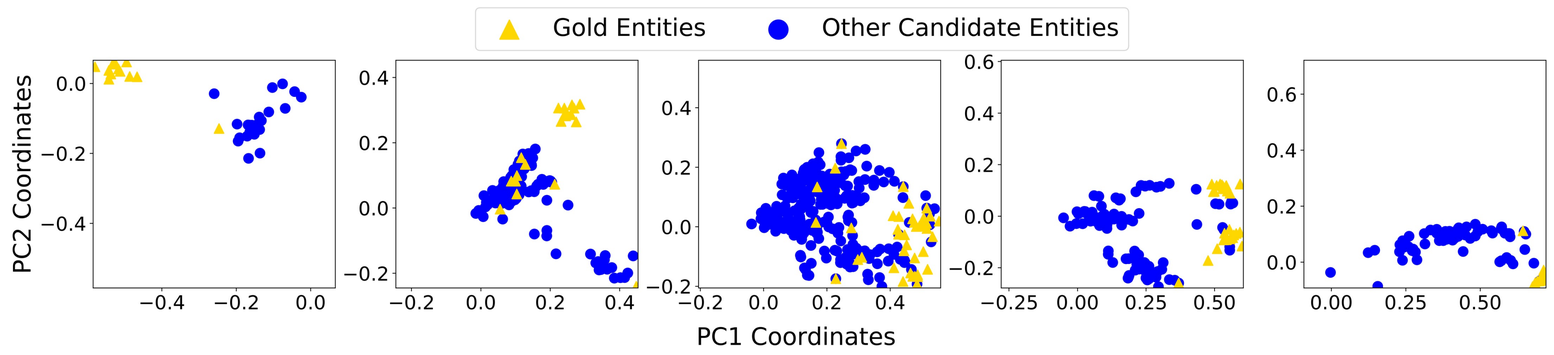
$\mathbf{E}_{m_i}$: Embeddings of candidate entities for mention m_i. One candidate embedding per row.
$\mathbf{E}_D$: Document embedding matrix for doc D. One candidate embedding per row.
$\mathcal{W}_{m_i}$: Weights of candidate entities for mention m_i.
$\mathbf{W}_D$: Weights of candidate entities for document D.



- ✓ Fully unsupervised
- ✓ Light-weight and scalable
- ✓ Explainable
- ✓ Language agnostic



Relationship between Eigenthemes and Gold Entities

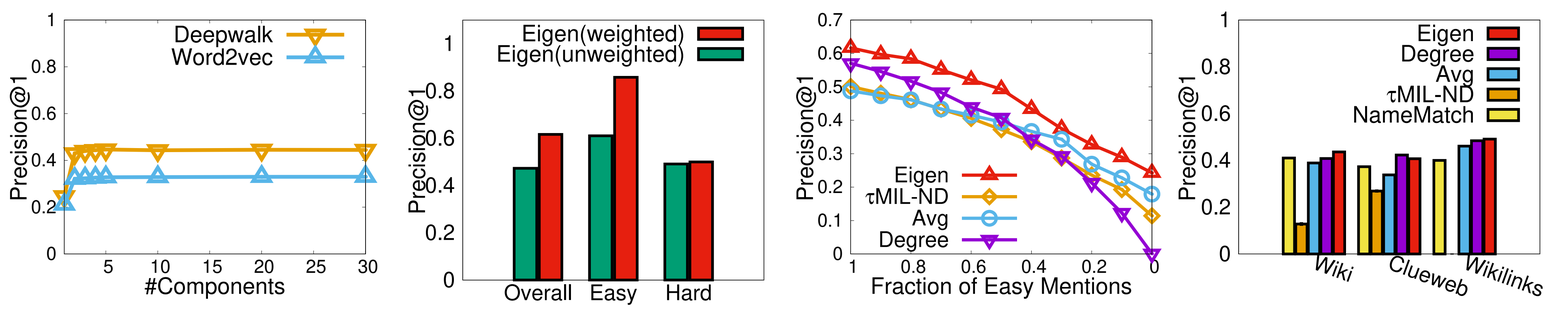


Results: CoNLL Dataset

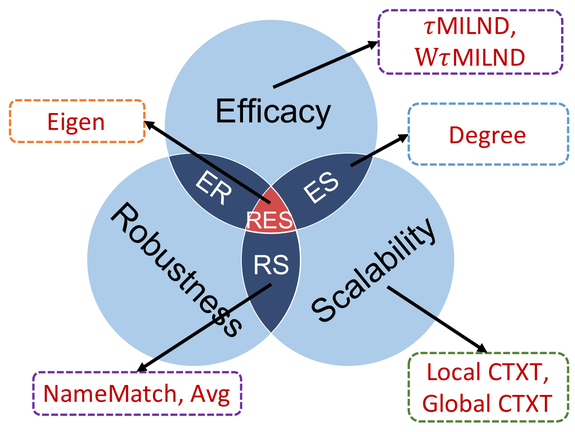
Category	Method	Precision@1			MRR		
		Overall#	Easy	Hard	Overall#	Easy	Hard
Existent	NAMEMATCH (Riedel et al., 2010)	0.412	0.645	0.174	0.415	0.645	0.184
Existent	τMIL-ND (SoTA) (Le and Titov, 2019)	0.451 ± 0.019	0.700 ± 0.032	0.187 ± 0.006	0.539 ± 0.017	0.777 ± 0.029	0.353 ± 0.005
Proposed	LOCAL CTXT	0.296	0.420	0.223	0.401	0.537	0.374
Proposed	GLOBAL CTXT	0.303	0.403	0.289	0.423	0.542	0.448
Proposed	DEGREE	0.571	**1.0**†	0.0	0.649	**1.0**†	0.312
Proposed	AVG	0.488	0.658	0.445	0.593	0.756	0.636
Proposed	WτMIL-ND	0.499 ± 0.022	0.778 ± 0.037	0.217 ± 0.008	0.592 ± 0.018	0.853 ± 0.030	0.415 ± 0.007
Proposed	EIGEN	**0.617**†	0.858	**0.500**†	**0.690**†	0.910	**0.674**†
-	Ceiling	0.824	1.0	1.0	0.824	1.0	1.0

† Indicates statistical significance ($p < 0.01$) between the best and the second-best method using the Student's paired t-test.
# Overall is computed by considering all mentions (including Not-found in addition to Easy and Hard).

Hyperparameter Tuning & Analysis



Summary



GitHub

PDF