Entity Linking via Low-rank Subspaces

Akhil Arora, Alberto García-Durán, and Bob West

SMLD

November 13, 2019

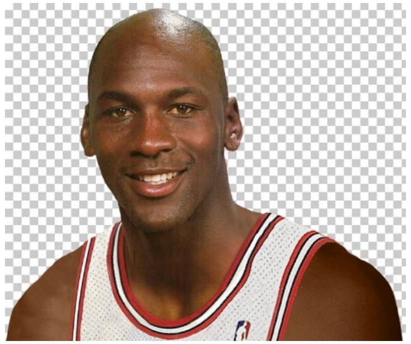
What is Entity Linking?

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

What is Entity Linking?

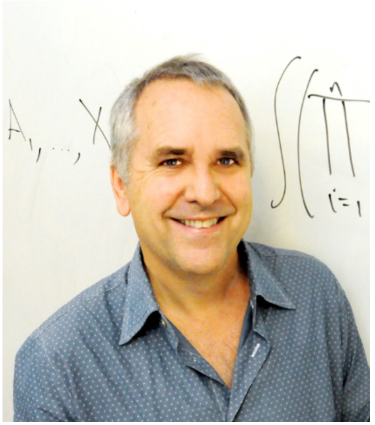
"Michael Jordan is one of the leading figures in machine learning, and in 2016 Science reported him as the world's most influential computer scientist."

What is Entity Linking?



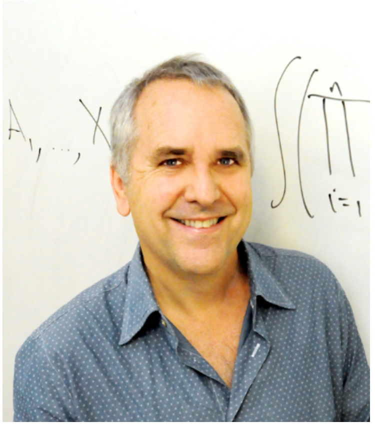
"Michael Jordan is one of the leading figures in machine learning, and in 2016 Science reported him as the world's most influential computer scientist."

What is Entity Linking?



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

What is Entity Linking?



en.wikipedia.org/wiki/Michael_I._Jordan

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

en.wikipedia.org/wiki/Science_(journal)

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps

"Michael Jordan"

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

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps

“Michael Jordan”

Candidate Entity	Prior P(e|m)
Michael_Jordan	0.997521
Michael	
Michael_Jc	
Michael_Jord	

Candidate Entity	Prior P(e|m)
Science	0.737955
Science_(journal)	0.207151
Science_Channel	0.005036

“Science”

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps
  - High quality candidate generation
  - Prior information: a strong feature

"Michael Jordan"

Candidate Entity	Prior P(e|m)
Michael_Jordan	0.997521
Michael_	
Michael_Jo	
Michael_Jord	

Candidate Entity	Prior P(e|m)
Science	0.737955
Science_(journal)	0.207151
Science_Channel	0.005036

"Science"

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps
  - High quality candidate generation
  - Prior information: a strong feature
- Other Features:
  - Local/Global context
  - Coherence in disambiguated entities

"Michael Jordan"

Candidate Entity	Prior P(e|m)
Michael_Jordan	0.997521
Michael_	
Michael_Jc	
Michael_Jord	

Candidate Entity	Prior P(e|m)
Science	0.737955
Science_(journal)	0.207151
Science_Channel	0.005036

"Science"

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps
  - High quality candidate generation
  - Prior information: a strong feature
- Other Features:
  - Local/Global context
  - Coherence in disambiguated entities
- Sophisticated Supervised Models
  - XGBoost
  - Deep Neural Networks

"Michael Jordan"

Candidate Entity	Prior P(e|m)
Michael_Jordan	0.997521
Michael_	
Michael_Jo	
Michael_Jord	

Candidate Entity	Prior P(e|m)
Science	0.737955
Science_(journal)	0.207151
Science_Channel	0.005036

"Science"

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps
  - High quality candidate generation
  - Prior information: a strong feature
- Other Features:
  - Local/Global context
  - Coherence in disambiguated entities
- Sophisticated Supervised Models
  - XGBoost
  - Deep Neural Networks

“Michael Jordan”

Candidate Entity	Prior P(e|m)
Michael_Jordan	0.997521
Michael_	
Michael_Jo	
Michael_Jord	

Candidate Entity	Prior P(e|m)
Science	0.737955
Science_(journal)	0.207151
Science_Channel	0.005036

“Science”

Sky is the limit ☺!

How to perform Entity Linking?

- Use Dictionaries/Alias-tables/Probability-Maps
  - High quality candidate generation
  - Prior information: a strong feature
- Other Features:
  - Local/Global context
  - Coherence in disambiguated entities
- Sophisticated Supervised Models
  - XGBoost
  - Deep Neural Networks

"Michael Jordan"

Candidate Entity	Prior P(e|m)
Michael_Jordan	0.997521
Michael	
Michael_Jc	
Michael_Jord	

Candidate Entity	Prior P(e|m)
Science	0.737955
Science_(journal)	0.207151
Science_Channel	0.005036

"Science"

Sky is the limit ☺!

[NAACL'18] SOTA P@1 = **95.9**

"NLP Progress: Entity Linking", http://nlpprogress.com/english/entity_linking.html

“Unaddressed” Research Questions

- Are dictionaries naturally available across use-cases?

“Unaddressed” Research Questions

- Are dictionaries naturally available across use-cases?
  - Lack of annotated data
    - Specialized Domains: Medical, Scientific, Legal, Enterprise specific corpora
  - Noisy and rapidly evolving annotated data
    - Web queries

"Unaddressed" Research Questions

- Are dictionaries naturally available across use-cases?
  - Lack of annotated data
    - Specialized Domains: Medical, Scientific, Legal, Enterprise specific corpora
  - Noisy and rapidly evolving annotated data
    - Web queries
- Can existing SOTA methods operate at Web Scale?

“Unaddressed” Research Questions

- Are dictionaries naturally available across use-cases?
  - Lack of annotated data
    - Specialized Domains: Medical, Scientific, Legal, Enterprise specific corpora
  - Noisy and rapidly evolving annotated data
    - Web queries
- Can existing SOTA methods operate at Web Scale?
  - We can only hope!

“Unaddressed” Research Questions

- Are dictionaries naturally available across use-cases?
  - Lack of annotated data
    - Specialized Domains: Medical, Scientific, Legal, Enterprise specific corpora
  - Noisy and rapidly evolving annotated data
    - Web queries
- Can existing SOTA methods operate at Web Scale?

hope!

"Unaddressed" Research Questions

- Are dictionaries naturally available across use-cases?
  - Lack of annotated data
    - Specialized Domains: Medical, Scientific, Legal, Enterprise specific corpora
  - Noisy and rapidly evolving annotated data
    - Web queries
- Can existing SOTA methods operate at Web Scale?

hope!

- NAACL'18 SOTA: 9 hours to train using 16 threads on CoNLL benchmark of only 18K entity mentions
- Some DL methods take more than 1 day

"Unaddressed" Research Questions

- Are dictionaries naturally available across use-cases?
  - Lack of annotated data
    - Specialized Domains: Medical, Scientific, Legal, Enterprise specific corpora
  - Noisy and rapidly evolving annotated data
    - Web queries
- Can existing SOTA methods operate at Web Scale?

hope!

- NAACL'18 SOTA: 9 hours to train using 16 threads on CoNLL benchmark of only 18K entity mentions
- Some DL methods take more than 1 day

Scalable EL without Annotated Data

Entity Linking without Annotated Data

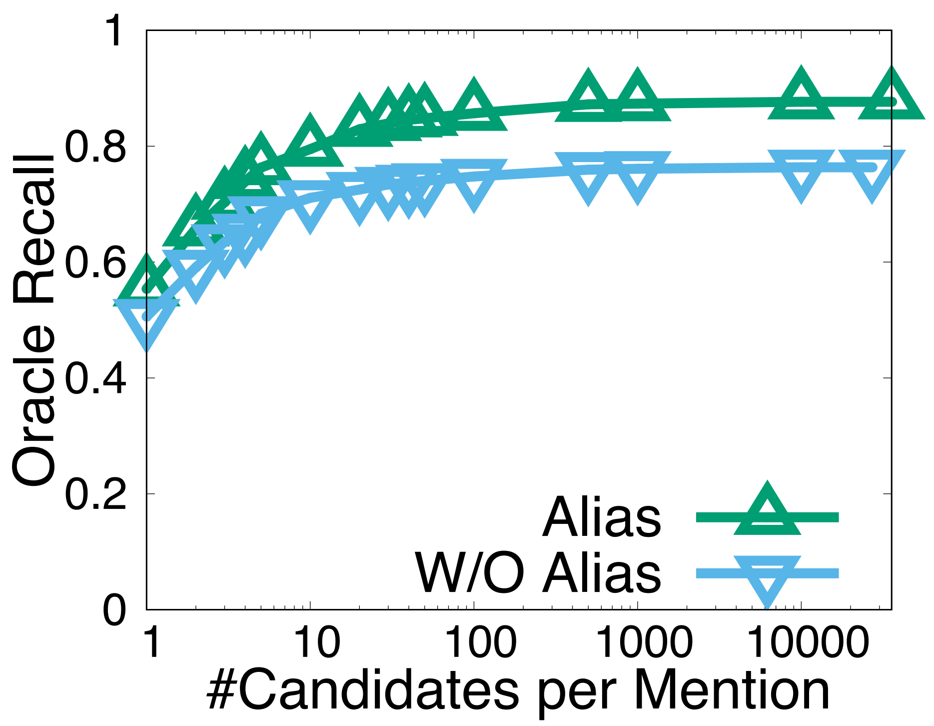
- Candidate generator
- Entity embeddings
  - Learn from the underlying graph
  - Learn from textual descriptions of entities
- Collective disambiguation
  - Ensures “topical coherence” among entities in a document

Candidate Generation

- Simple yet practical
  - Candidates contain all tokens of the mention
  - Example: For mention “Michael Jordan”
    - Michael Jordan (basketball player) and Michael Jordan (computer scientist) are candidates
    - Michael Jackson is not
  - Rank candidates using entity degree (relates to popularity)

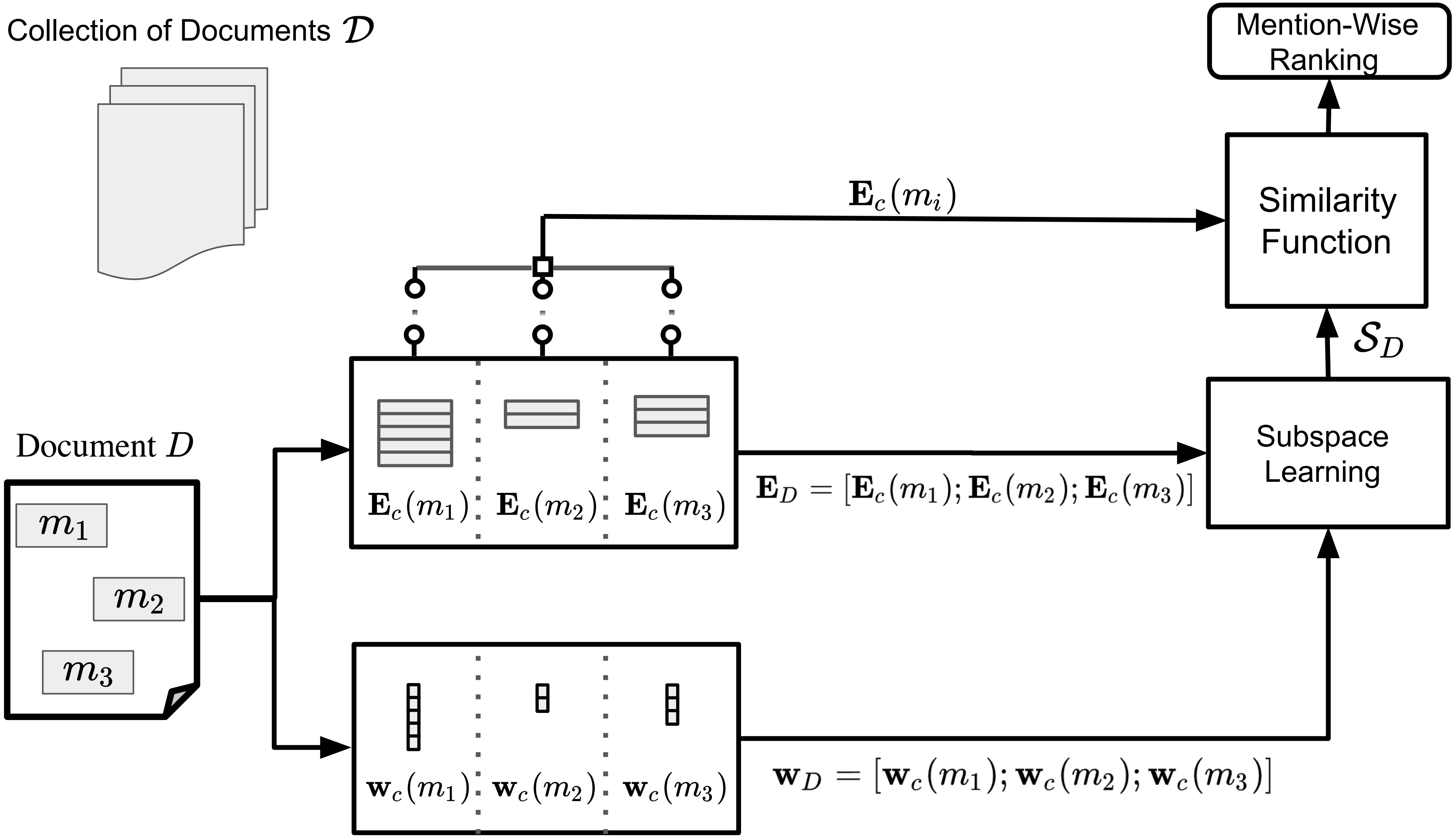
Candidate Generation

- Simple yet practical
  - Candidates contain all tokens of the mention
  - Example: For mention “Michael Jordan”
    - Michael Jordan (basketball player) and Michael Jordan (computer scientist) are candidates
    - Michael Jackson is not
  - Rank candidates using entity degree (relates to popularity)



- Aliases of entity names to boost recall

Eigenthemes for Entity Disambiguation



Subspace Learning: Intuition

Subspace captures the main "theme" of a document

"Science"

Candidate Entity
Science
Science_(journal)
Science_Channel

"Michael Jordan"

Candidate Entity
Michael_Jordan
Michael_I._Jordan
Michael_Jordan_statue
Michael_Jordan_(footballer)

Subspace Learning: Intuition

Subspace captures the main "theme" of a document

"Science"

Candidate Entity
Science
Science_(journal)
Science_Channel

"Michael Jordan"

Candidate Entity
Michael_Jordan
Michael_I._Jordan
Michael_Jordan_statue
Michael_Jordan_(footballer)

Top-k d-dimensional eigen vectors of the covariance matrix of candidate entity embeddings in a document

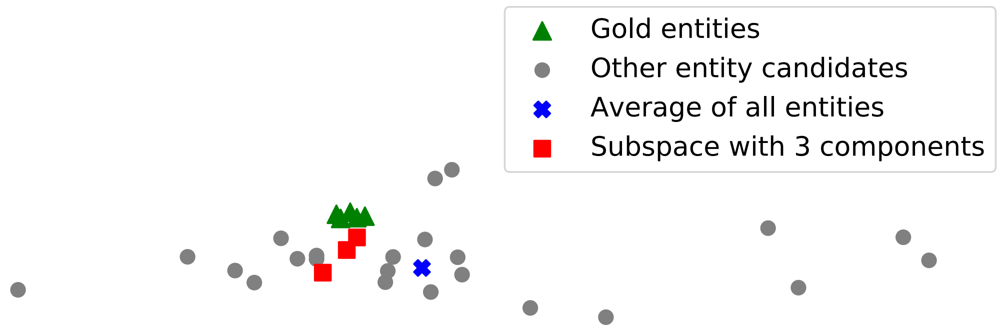
Subspace Learning: Intuition

Subspace captures the main “theme” of a document

“Science”
Candidate Entity
Science
Science_(journal)
Science_Channel

“Michael Jordan”
Candidate Entity
Michael_Jordan
Michael_I._Jordan
Michael_Jordan_statue
Michael_Jordan_(footballer)

Top-k d-dimensional eigen vectors of the covariance matrix of candidate entity embeddings in a document



Subspace Learning: Intuition

Subspace captures the main "theme" of a document

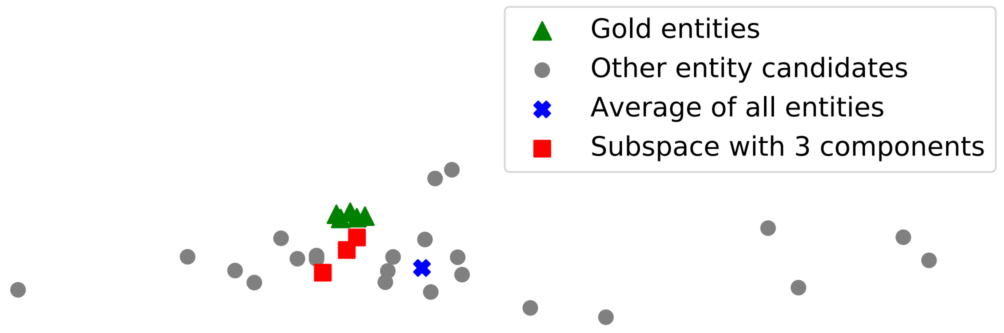
"Science"

Candidate Entity
Science
Science_(journal)
Science_Channel

"Michael Jordan"

Candidate Entity
Michael_Jordan
Michael_I._Jordan
Michael_Jordan_statue
Michael_Jordan_(footballer)

Top-k d-dimensional eigen vectors of the covariance matrix of candidate entity embeddings in a document



External signals to enrich subspace learning

– Eigendecomposition of the weighted covariance matrix

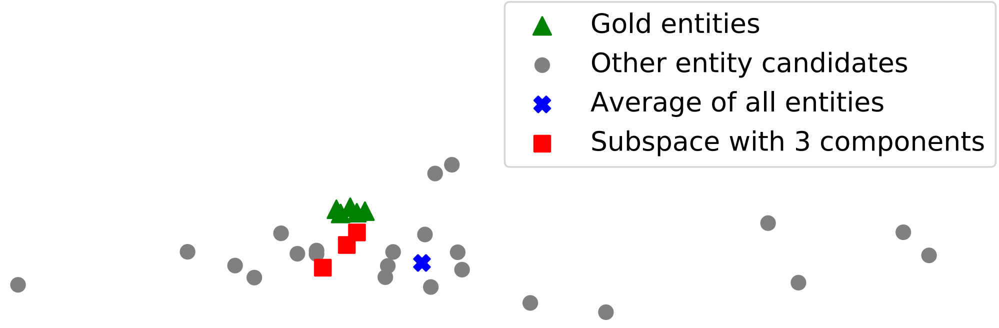
Subspace Learning: Intuition

Subspace captures the main “theme” of a document

“Science”
Candidate Entity
Science
Science_(journal)
Science_Channel

“Michael Jordan”
Candidate Entity
Michael_Jordan
Michael_I._Jordan
Michael_Jordan_statue
Michael_Jordan_(footballer)

Top-k d-dimensional eigen vectors of the covariance matrix of candidate entity embeddings in a document



External signals to enrich subspace learning

- Eigendecomposition of the weighted covariance matrix
- Entity embeddings with high weights act as “anchor embeddings”
  - Prioritized in subspace learning
- Weighting scheme: Inverse of the rank computed using entity degree information

Setup

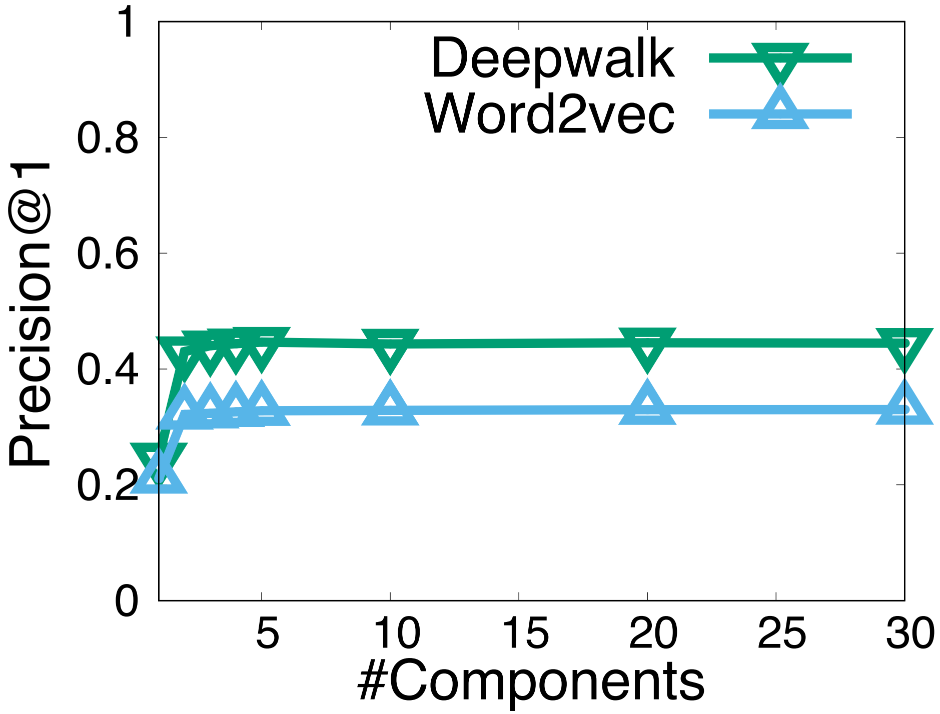
- Datasets
  - **CoNLL:** Most popular benchmark dataset for EL, based on CoNLL 2003 shared task
  - **More in the Paper:**
    - **WNED (Wiki and Clueweb):** Benchmarks from English Wikipedia and Clueweb corpora
    - **Wikilinks-Random:** Tables extracted from English Wikipedia
- Referent KB: Wikidata

Setup

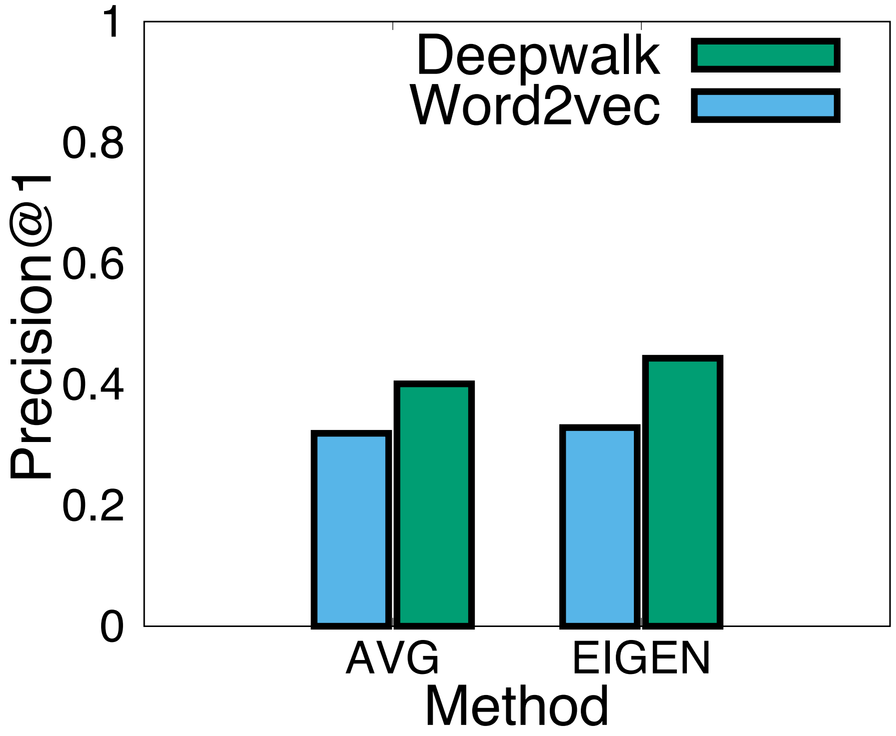
- Datasets
  - **CoNLL:** Most popular benchmark dataset for EL, based on CoNLL 2003 shared task
  - **More in the Paper:**
    - **WNED (Wiki and Clueweb):** Benchmarks from English Wikipedia and Clueweb corpora
    - **Wikilinks-Random:** Tables extracted from English Wikipedia
- Referent KB: Wikidata
- Embeddings:
  - Words: Pre-trained Word2vec
  - Entity embeddings:
    - Deepwalk trained on Wikidata
    - Average of Word2vec vectors of entity description words

Tuning on CoNLL-Val

Tuning #components



Impact of entity embedding technique on EL



Baselines

- **NameMatch:**
  - Retrieves all entities whose names match exactly with the mention string
  - Ties are broken using entity degree

Baselines

- **NameMatch:**
  - Retrieves all entities whose names match exactly with the mention string
  - Ties are broken using entity degree
- **Degree:**
  - Candidates are ranked based on entity degree
  - Highest degree candidate entity is the prediction for a given mention
- **Avg and WAvg:**
  - (Weighted)Avg of candidate embeddings in a document as its representation
  - Most similar candidate (Cosine Sim) with the doc representation is the prediction

Baselines

- **NameMatch:**
  - Retrieves all entities whose names match exactly with the mention string
  - Ties are broken using entity degree
- **Degree:**
  - Candidates are ranked based on entity degree
  - Highest degree candidate entity is the prediction for a given mention
- **Avg and WAvg:**
  - (Weighted)Avg of candidate embeddings in a document as its representation
  - Most similar candidate (Cosine Sim) with the doc representation is the prediction
- **Le and Titov:** Uses weak supervision or distant learning
  - Candidate entities of a mention (which might miss the 'true' entity) are scored higher than a number of randomly sampled entities
  - Rank based on similarity between candidates and the mention context

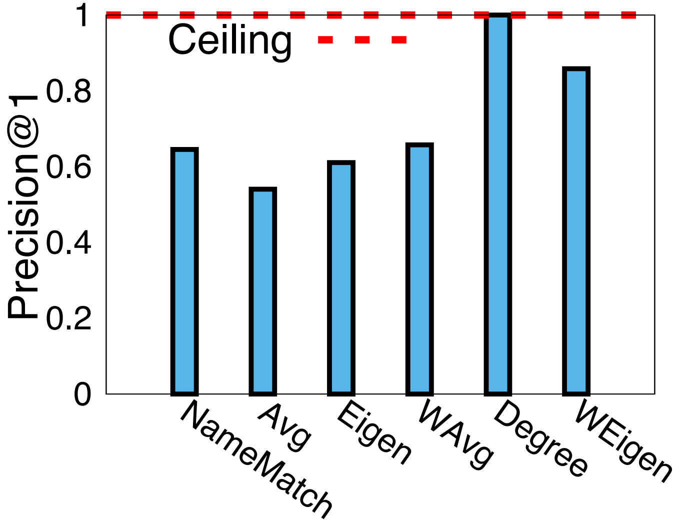
Is Eigenthemes Effective?

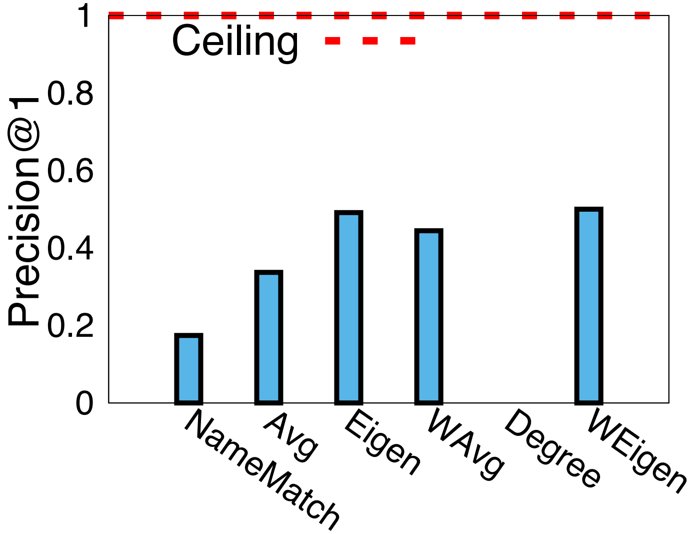
Dataset	Precision@1						
	NameMatch	Avg	Eigen	WAvg	Degree	WEigen	Ceiling
CoNLL-Test	0.412	0.394	0.473	0.488	0.571	**0.617**	0.824

Is Eigenthemes Effective?

Dataset	Precision@1						
	NameMatch	Avg	Eigen	WAvg	Degree	WEigen	Ceiling
CoNLL-Test	0.412	0.394	0.473	0.488	0.571	**0.617**	0.824

Easy Mentions: Degree ranks gold entity at the top





Hard Mentions: Gold entity not at the top using degree

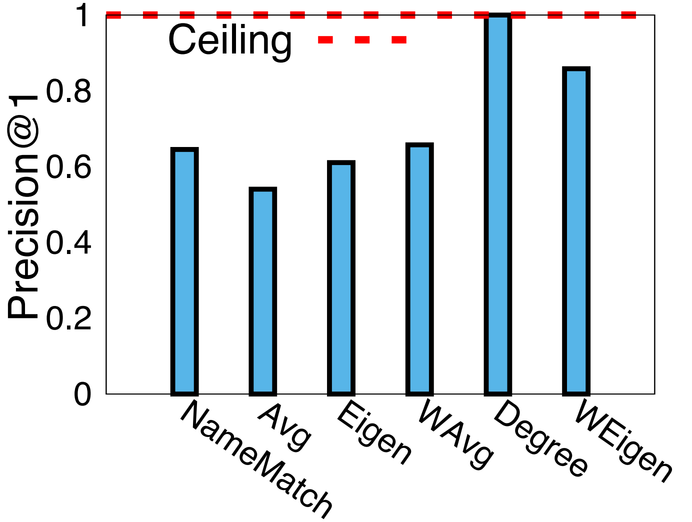
Is Eigenthemes Effective?

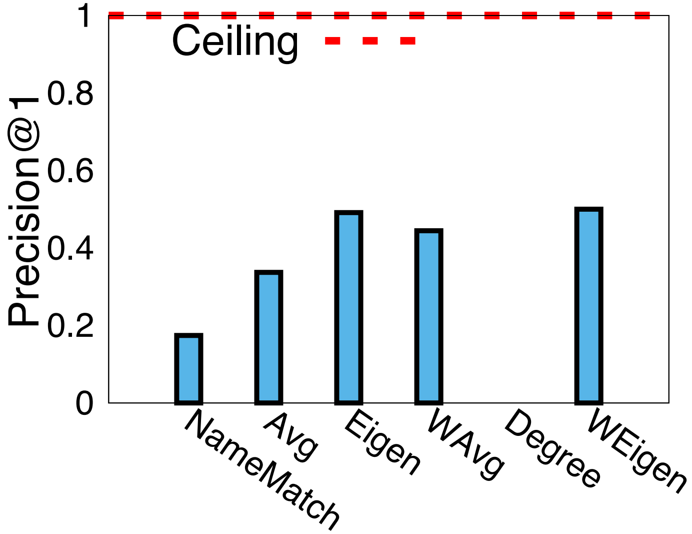
Dataset	Precision@1						
	NameMatch	Avg	Eigen	WAvg	Degree	WEigen	Ceiling
CoNLL-Test	0.412	0.394	0.473	0.488	0.571	**0.617**	0.824

Precision@1 in Le and Titov's CoNLL Test Dataset

Technique	NameMatch	τMIL-ND	Freebase Prominence	Degree
Le and Titov's implementation[21]	0.150	0.389	-	-
Our Implementation	0.299	NA	0.326	**0.399**

Easy Mentions: Degree ranks gold entity at the top





Hard Mentions: Gold entity not at the top using degree

Is Eigenthemes Effective?

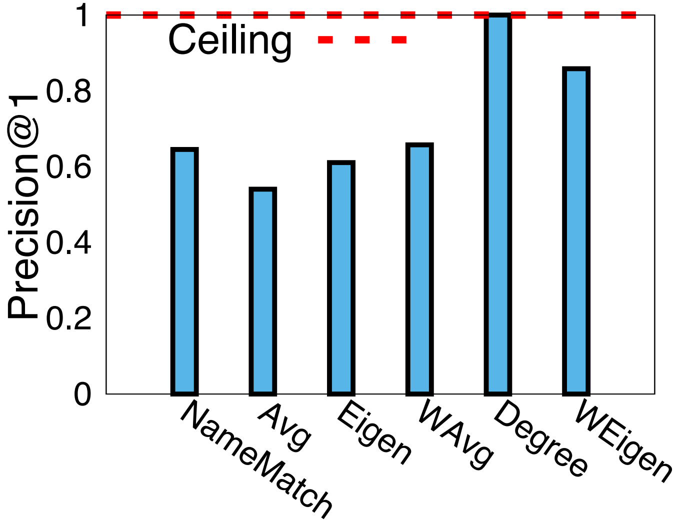
Dataset	Precision@1						
	NameMatch	Avg	Eigen	WAvg	Degree	WEigen	Ceiling
CoNLL-Test	0.412	0.394	0.473	0.488	0.571	**0.617**	0.824

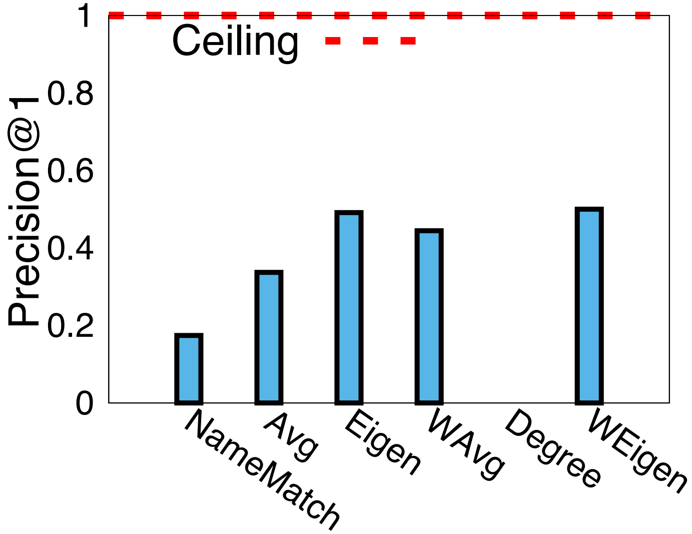
Precision@1 in Le and Titov's CoNLL Test Dataset

Technique	NameMatch	τMIL-ND	Freebase Prominence	Degree
Le and Titov's implementation[21]	0.150	0.389	-	-
Our Implementation	0.299	NA	0.326	**0.399**

Using Eigenthemes score as a feature for Supervised models portrays significant performance improvements

Easy Mentions: Degree ranks gold entity at the top





Hard Mentions: Gold entity not at the top using degree

Takeaways

👍 A single hyperparameter (#components) – ease of tuning for unannotated data

👍 Light-weight and scalable

– < 10 min for CoNLL, approx. 20 times faster than existing SOTA

👍 Language independence

👍 Ability to incorporate external signals as weights

Takeaways

- 👍 A single hyperparameter (#components) – ease of tuning for unannotated data
- 👍 Light-weight and scalable
  - – < 10 min for CoNLL, approx. 20 times faster than existing SOTA
- 👍 Language independence
- 👍 Ability to incorporate external signals as weights
- 👎 Early work that just scratches the surface

Takeaways

👍 A single hyperparameter (#components) – ease of tuning for unannotated data

👍 Light-weight and scalable

– < 10 min for CoNLL, approx. 20 times faster than existing SOTA

👍 Language independence

👍 Ability to incorporate external signals as weights

👎 Early work that just scratches the surface

– Candidate generation too simplistic
– Quality of entity embeddings can be improved
– Other tricks to boost performance …

THANK YOU

Questions?