RESLVE: Leveraging User Interest to Improve Entity Disambiguation on Short Text

Elizabeth L. Murnane elm236@cornell.edu
Bernhard Haslhofer bernhard.haslhofer@univie.ac.at
Carl Lagoze clagoze@umich.edu

Cornell University

A Personalized Approach to Entity Resolution

Background

- Task Definitions
- Challenges & Examples
- Attempted Solutions

Approach

- Motivations
- Modeling a Knowledge Context
- Implementation: The RESLVE System

Evaluation

- Experiments
- Results
- Future Work

A Personalized Approach to Entity Resolution

Background

- Task Definitions
- Challenges & Examples
- Attempted Solutions

Approach

- Motivations
- Modeling a Knowledge Context
- Implementation: The RESLVE System

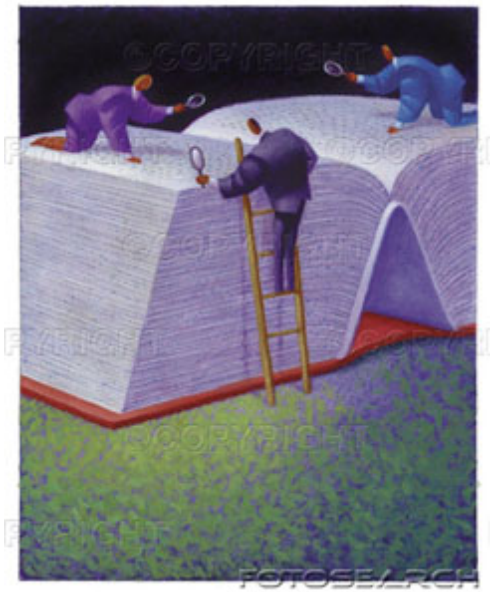
Evaluation

- Experiments
- Results
- Future Work


You
Tube

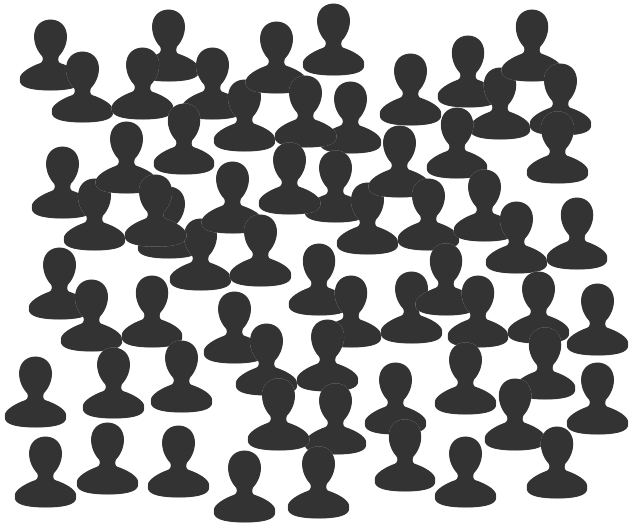
Social Web

10 million
pages per day



Social Web

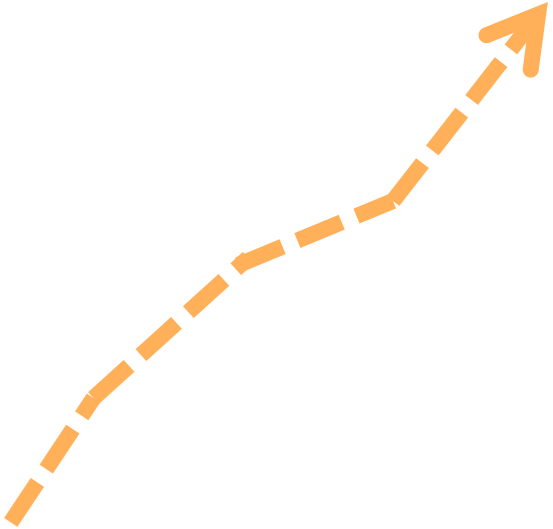
800 million
visitors per month



Social Web

flickr

7 billion images
(twice 4 years ago)



Task Definition

Task Definition

Named Entity Recognition (NER)

- Systematically identifying mentions of ***entities*** (e.g., people, places, concepts, ideas)

Task Definition

Named Entity Recognition (NER)

- Systematically identifying mentions of ***entities*** (e.g., people, places, concepts, ideas)

Named Entity Disambiguation (NED)

Resolving the intended meaning of ambiguous entities from multiple ***candidate meanings***

Ambiguous Entities

*aaahh one more day until **finn!!!** #cantwait*

***office** holiday party*

Beetle

Ambiguous Entities

aaahh one more day until finn!!! #cantwait

office holiday party

Beetle

Ambiguous Entities

aaahh one more day until finn!!! #cantwait

office holiday party

Beetle

Ambiguous Entities

aaahh one more day until finn!!! #cantwait

office holiday party

Beetle

Footage:

Footage:

- **Workplace?**

office holiday party

Footage:

- **Workplace?**
- **TV Show?**

office holiday party

Episode 4

Footage:

- ~~**Workplace?**~~
- **TV Show?**

office holiday party

Episode 4

Footage:

- ~~**Workplace?**~~
- **TV Show?**
  - **US Version?**
  - **UK Version?**

Episode 4

office, december 3

Footage:

- ~~**Workplace?**~~
- **TV Show?**
  - **US Version?**
  - **UK Version?**

Challenges & Focus

Challenges & Focus

- Short Length

Challenges & Focus

- Short Length
- Sparse Lexical Context

Challenges & Focus

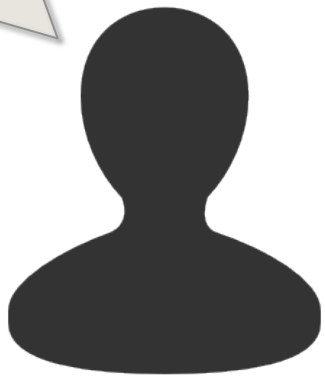
- Short Length
- Sparse Lexical Context
- Noisy

Challenges & Focus

- Short Length
- Sparse Lexical Context
- Noisy
- Highly personal in nature

Challenges & Focus

- Short Length
- Sparse Lexical Context
- Noisy
- Highly personal in nature



Limitations of Extant Research

Tweets severely degrade traditional techniques

Limitations of Extant Research

<u>Tweets severely degrade traditional techniques</u>

- Stanford NER: F1 drops 90% → 46%
- DBPedia Spotlight & Wikipedia Miner: P@1 < 40%

Limitations of Extant Research

Tweets severely degrade traditional techniques

- Stanford NER: F1 drops 90% → 46%
- DBPedia Spotlight & Wikipedia Miner: P@1 < 40%

Recent strategies

Limitations of Extant Research

<u>Tweets severely degrade traditional techniques</u>

- Stanford NER: F1 drops 90% → 46%
- DBPedia Spotlight & Wikipedia Miner: P@1 < 40%

<u>Recent strategies</u>

- Crowd-sourcing
  - Limitation: Dependent on reliable human workers

Limitations of Extant Research

<u>Tweets severely degrade traditional techniques</u>

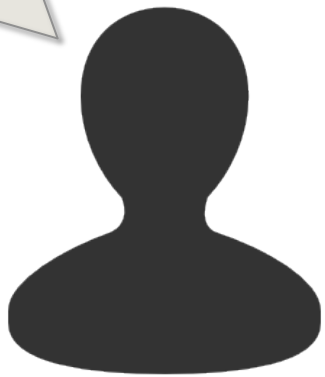
- Stanford NER: F1 drops 90% → 46%
- DBPedia Spotlight & Wikipedia Miner: P@1 < 40%

<u>Recent strategies</u>

- Crowd-sourcing
  - Limitation: Dependent on reliable human workers
- Automated attempts
  - Limitation: Focus on NER not NED
  - Limitation: Generalizability beyond Twitter?

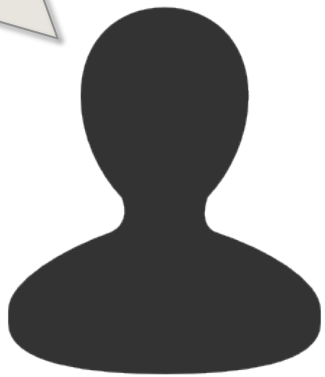
Challenges & Focus

- Short Length
- Sparse Lexical Context
- Noisy
- Highly personal in nature



Challenges & Focus

- Short Length
- Sparse Lexical Context
- Noisy
- Highly personal in nature



- User's past content on same platform not feasible background corpus

Task Definition

Named Entity Recognition (NER)

- Systematically identifying mentions of ***entities*** (e.g., people, places, concepts, ideas)

Named Entity Disambiguation (NED)

Resolving the intended meaning of ambiguous entities from multiple ***candidate meanings***

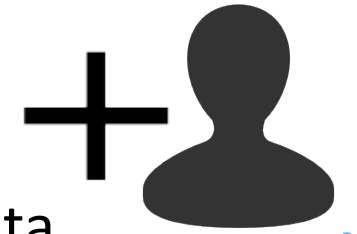
Our focus: disambiguating any entity detected in users' text-based utterances on social Web

Exploring a Personalized Solution

- Individual-centric approach to NED

Exploring a Personalized Solution

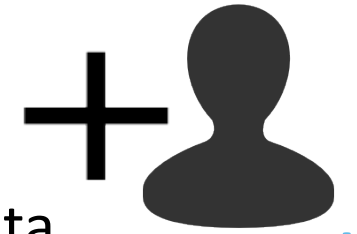
- Individual-centric approach to NED
- Incorporates external, user-specific semantic data



Personal Context

Exploring a Personalized Solution

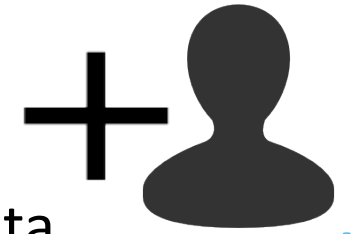
- Individual-centric approach to NED
- Incorporates external, user-specific semantic data
- Model personal interests with respect to this information



Personal Context

Exploring a Personalized Solution

- Individual-centric approach to NED
- Incorporates external, user-specific semantic data

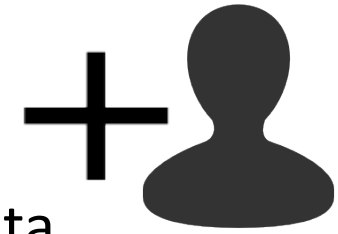


Personal Context

- Model personal interests with respect to this information
- Determine user's likely intended meaning of ambiguous entity based on similarity between potential meanings and interests

Exploring a Personalized Solution

- Individual-centric approach to NED
- Incorporates external, user-specific semantic data
- Model personal interests with respect to this information
- Determine user's likely intended meaning of ambiguous entity based on similarity between potential meanings and interests



Personal Context

RESLVE

Resolving **E**ntity **S**ense by **L**eVeraging **E**dits

Agenda

Background

- Task Definitions
- Challenges & Examples
- Attempted Solutions

Approach

- Motivations
- Modeling a Knowledge Context
- Implementation: The RESLVE System

Evaluation

- Experiments
- Results
- Future Work

Underlying Assumptions

Underlying Assumptions

- User has core interests
  - User more likely to mention an entity about a topic relevant to personal interests than mention a topic of non-interest

Underlying Assumptions

- User has core interests
  - User more likely to mention an entity about a topic relevant to personal interests than mention a topic of non-interest
- User expresses these interests consistently in content she posts online in multiple communities

Underlying Assumptions

- User has core interests
  - User more likely to mention an entity about a topic relevant to personal interests than mention a topic of non-interest
- User expresses these interests consistently in content she posts online in multiple communities
- Can use a semantic knowledge base to formally represent these topics of interest

Underlying Assumptions

- User has core interests
  - User more likely to mention an entity about a topic relevant to personal interests than mention a topic of non-interest
- User expresses these interests consistently in content she posts online in multiple communities
- Can use a semantic knowledge base to formally represent these topics of interest

➢ Bridge user identity between social Web and knowledge base, K
➢ Model interests using K's organizational scheme
➢ Rank entity senses according to relevance to interests

Qualitative Analysis: Stable Interests

Qualitative Analysis: Stable Interests

User's topics of contribution similar across Web:

Qualitative Analysis: Stable Interests

User's topics of contribution similar across Web:

<u>Same Topics</u>

Ambiguous YouTube post:
office, december 3

Same user's recent Wikipedia edit:
<item userid="xxxx" user="xxxx" pageid="31841130" title= "The Office (U.S. season 8)"/>

Qualitative Analysis: Stable Interests

User's topics of contribution similar across Web:

Same Topics

Ambiguous YouTube post:
office, december 3

Same user's recent Wikipedia edit:
<item userid="xxxx" user="xxxx" pageid="31841130" title= "The Office (U.S. season 8)"/>

Same categories

- On average, 52.4% of entities a user mentions in social Web (e.g., "Java") have at least 1 candidate sense in same parent category of Wikipedia article same user edited (e.g., "Programming language")
- If extend to just 4 parents up category hierarchy, get all 100%

Theoretical Motivations

Theoretical Motivations

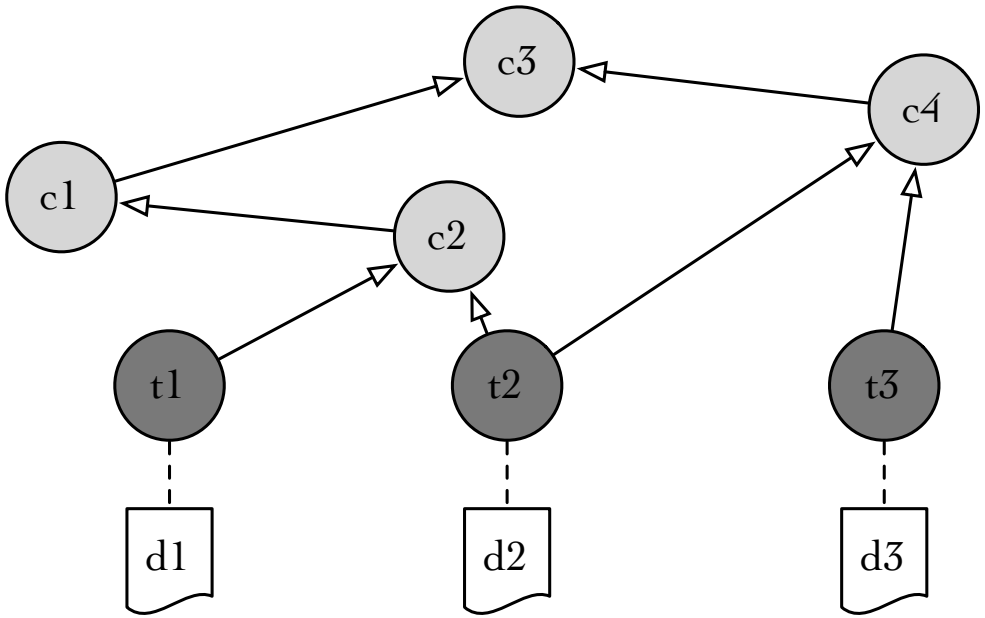
- Online Contribution:
  - Users produce online content about key set of personally-interesting topics because it is fulfilling and seen as having better cost benefit
  - (Harper et al., 2007; Lakhani & von Hippel, 2003; Lerner & Tirole, 2000; Ling et al., 2006; Maslow, 1970)

Theoretical Motivations

- Online Contribution:
  - Users produce online content about key set of personally-interesting topics because it is fulfilling and seen as having better cost benefit
  - (Harper et al., 2007; Lakhani & von Hippel, 2003; Lerner & Tirole, 2000; Ling et al., 2006; Maslow, 1970)
- Modeling Interests:
  - Effective to model these topic interests from lexical features of these text-based contributions
  - (Chen et al., 2010; Cosley et al., 2007; Pennacchiotti & Popescu, 2011)

Modeling a Knowledge Context

- Knowledge base, *K*
- *K=(N,E)*
- 2 node types:
  - Categories
  - Topics



The Knowledge Graph

The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$

The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$

The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$

The Knowledge Graph

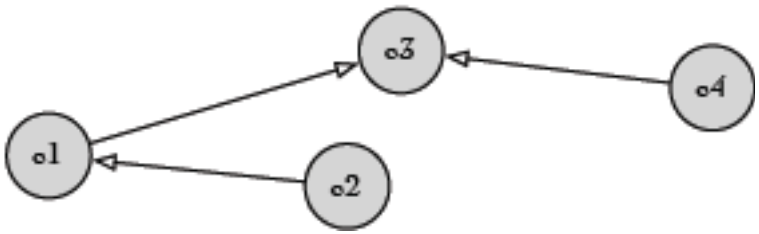
- **Category** nodes: $N_{Category} \subset N$

The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier

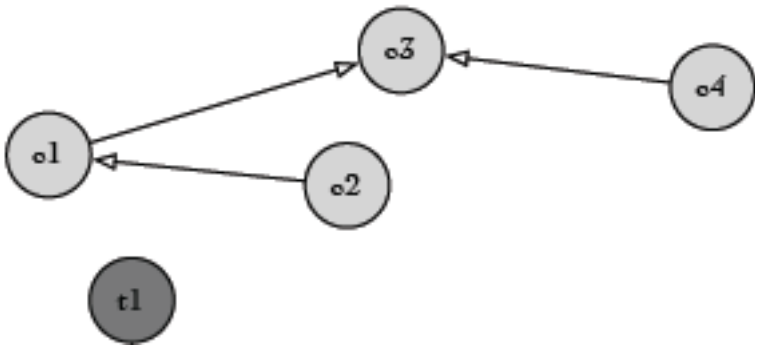
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes



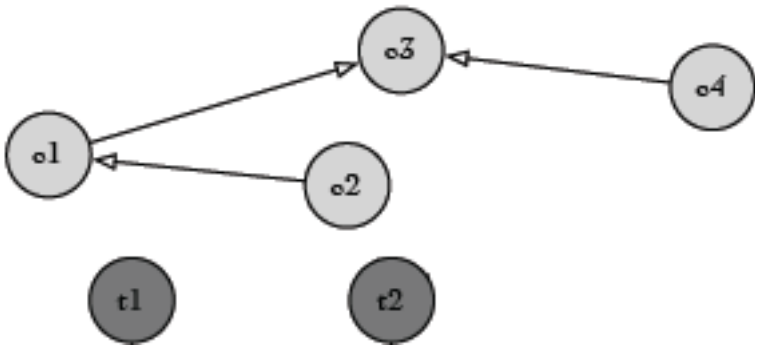
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$



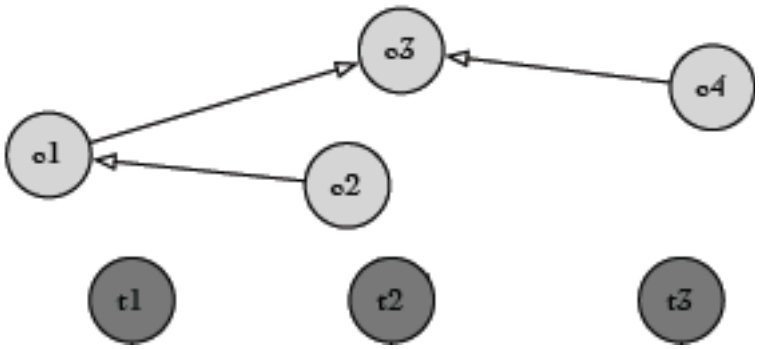
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$



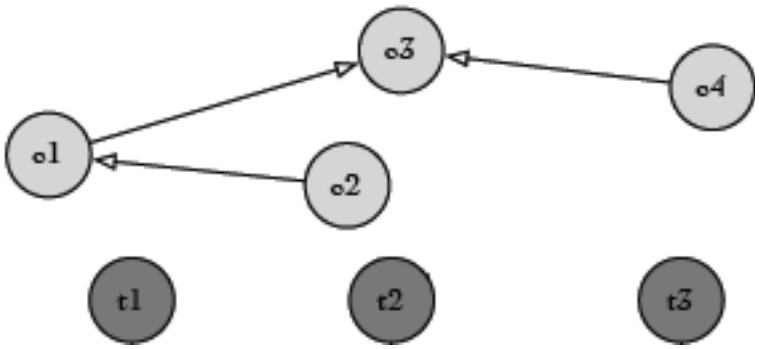
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$



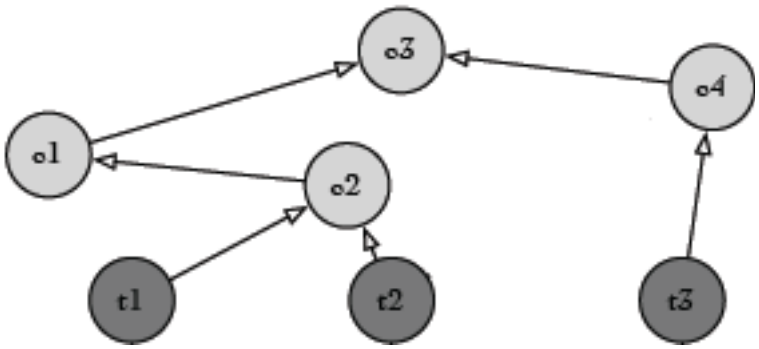
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier



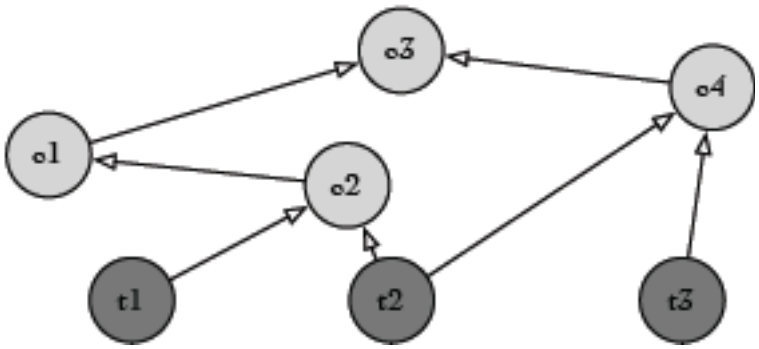
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier



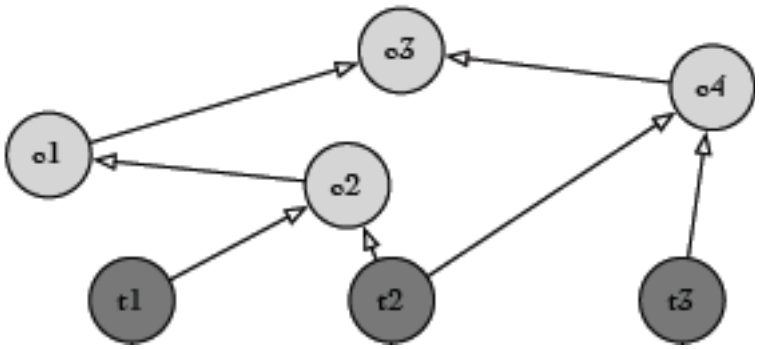
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier



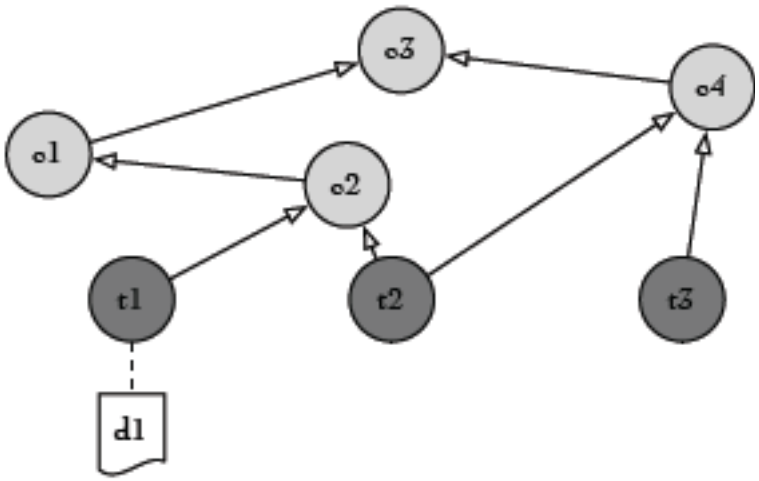
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier
  - Belongs to one or more categories



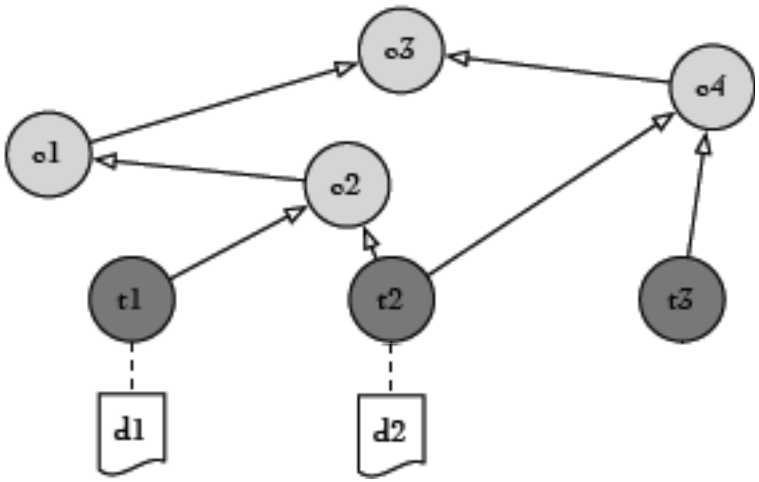
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier
  - Belongs to one or more categories



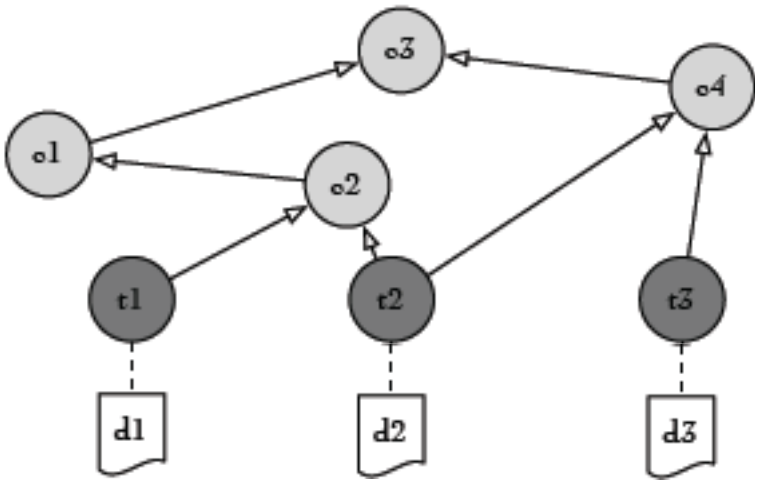
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier
  - Belongs to one or more categories



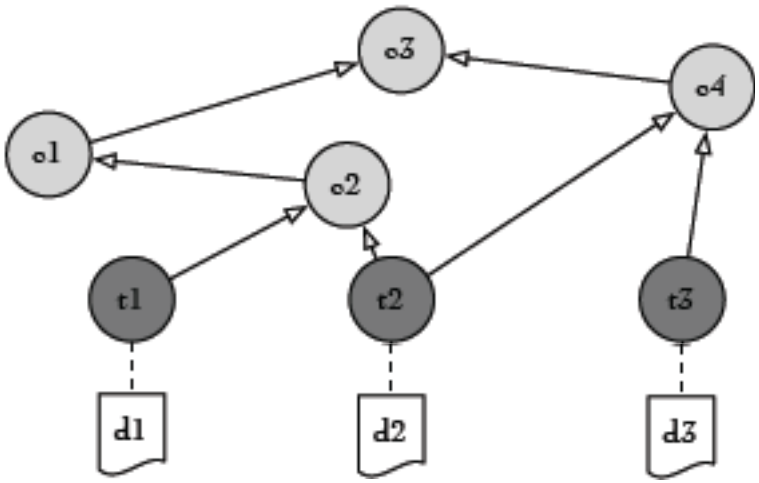
The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier
  - Belongs to one or more categories



The Knowledge Graph

- **Category** nodes: $N_{Category} \subset N$
  - Unique identifier
  - Semantic relationships with other nodes
- **Topic** nodes: $N_{Topic} \subset N$
  - Unique identifier
  - Belongs to one or more categories
  - Associated with text-based description



User Interest Model

User Interest Model

- Editing a description signals interest in associated topic

User Interest Model

- Editing a description signals interest in associated topic

User Interest Model

- Editing a description signals interest in associated topic

User Interest Model

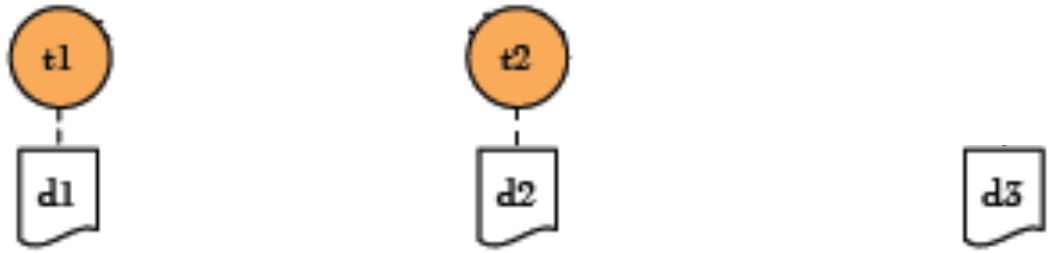
- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of

User Interest Model

- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of

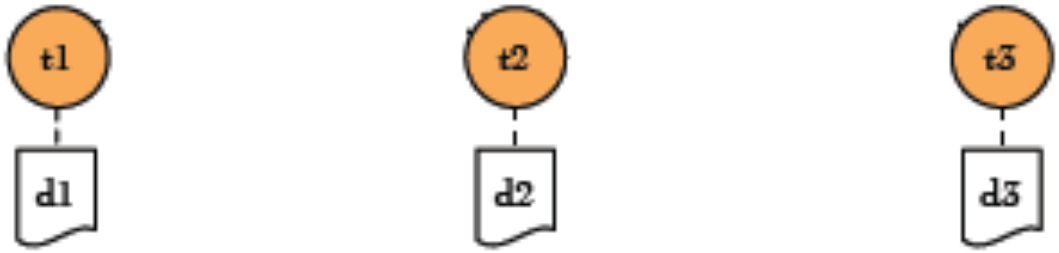
User Interest Model

- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of



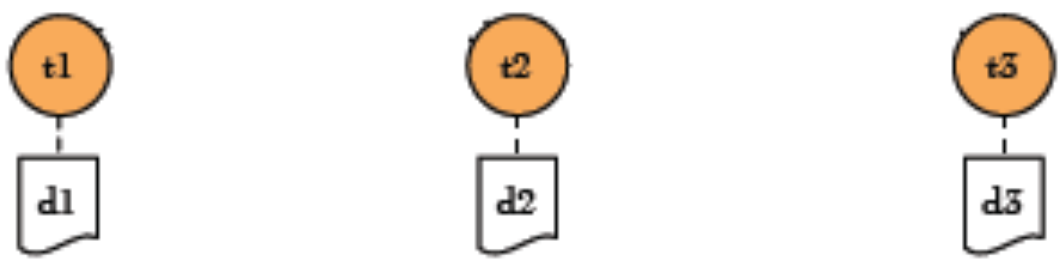
User Interest Model

- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of



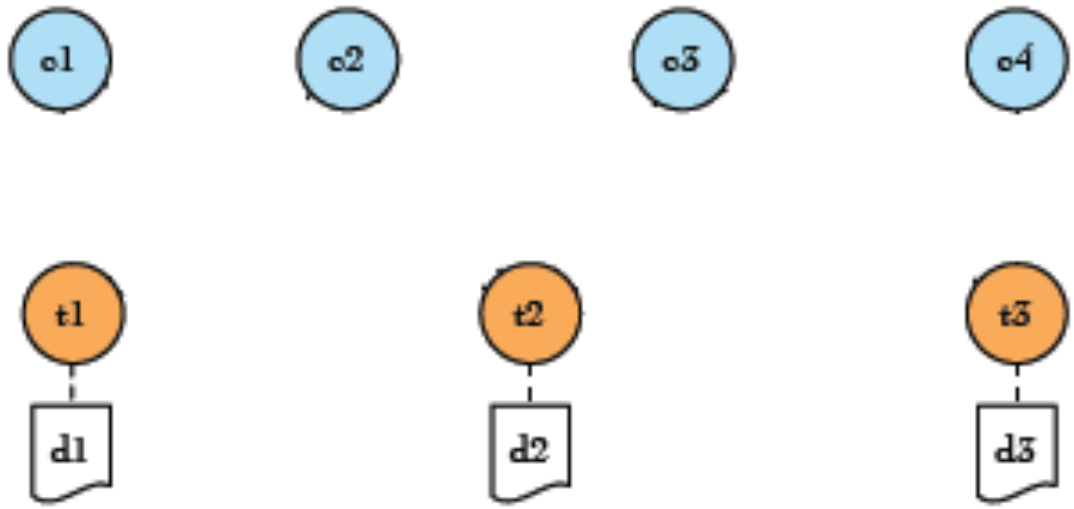
User Interest Model

- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of
- Category nodes: categories reachable in knowledge graph from those topics



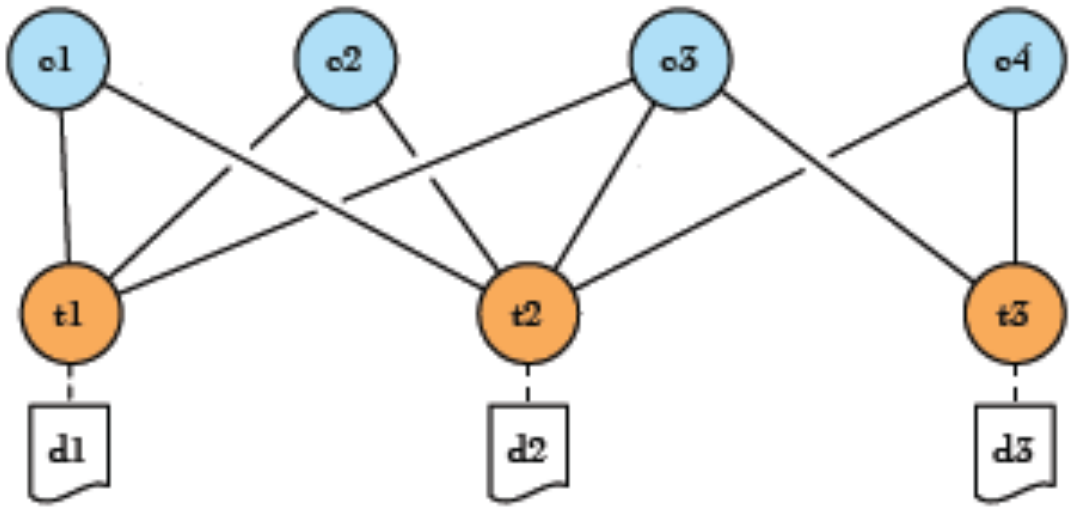
User Interest Model

- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of
- Category nodes: categories reachable in knowledge graph from those topics



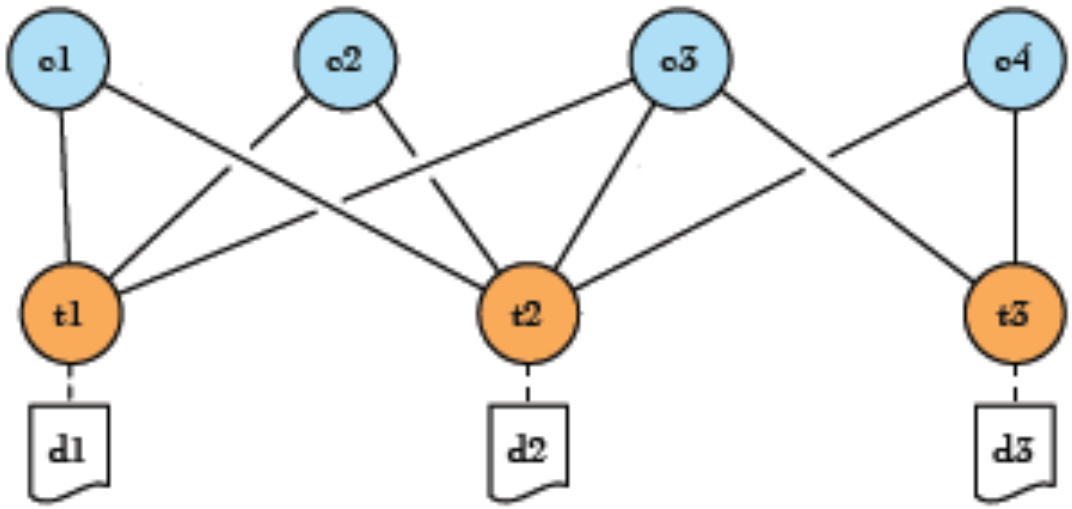
User Interest Model

- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of
- Category nodes: categories reachable in knowledge graph from those topics



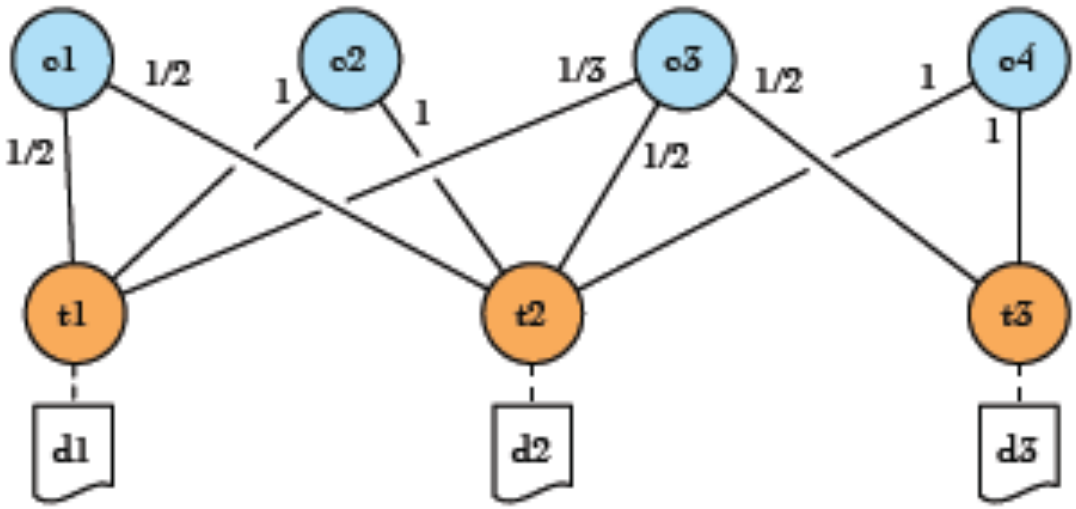
User Interest Model

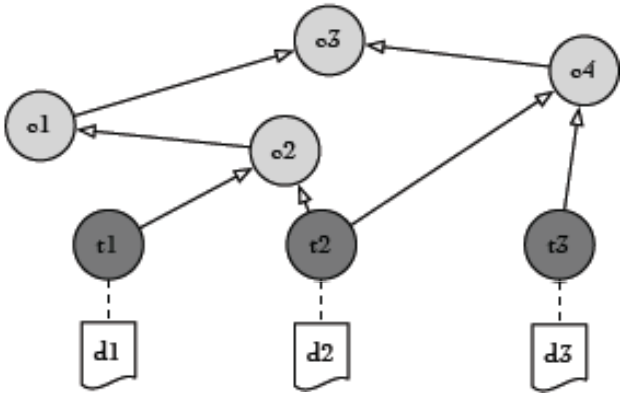
- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of
- Category nodes: categories reachable in knowledge graph from those topics
- Edge weight = inverse of shortest path length



User Interest Model

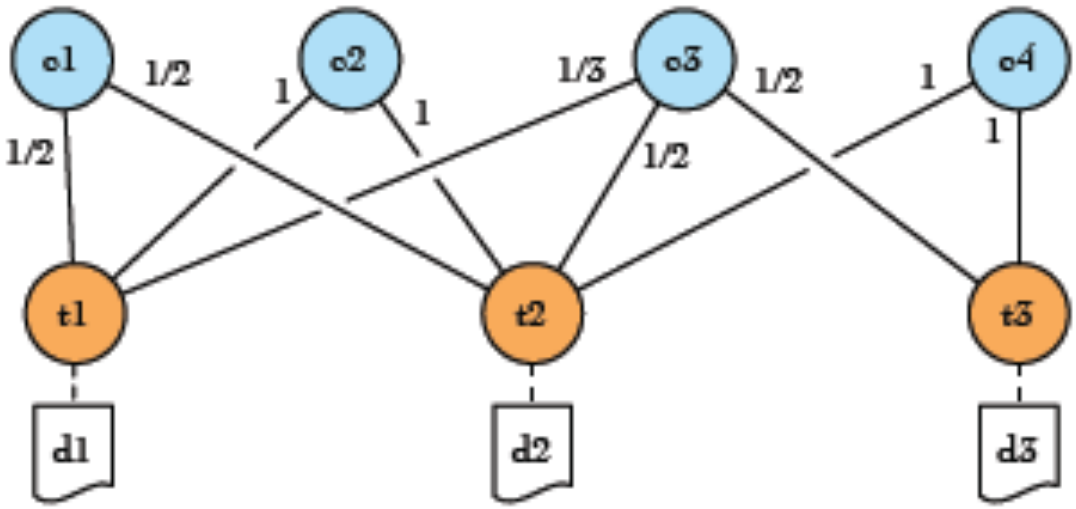
- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of
- Category nodes: categories reachable in knowledge graph from those topics
- Edge weight = inverse of shortest path length





User Interest Model

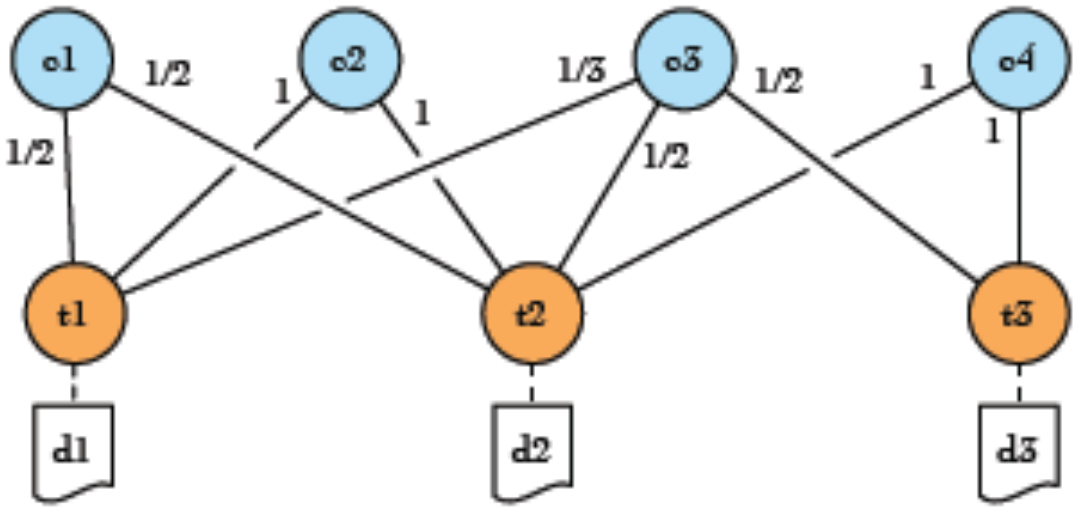
- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of
- Category nodes: categories reachable in knowledge graph from those topics
- Edge weight = inverse of shortest path length



	c1	*c2*	*c3*	*c4*
t1	$\frac{1}{2}$	1	$\frac{1}{3}$	0
t2	$\frac{1}{2}$	1	$\frac{1}{2}$	1
t3	0	0	$\frac{1}{2}$	1

User Interest Model

- Editing a description signals interest in associated topic
- Topic nodes: all topics user edited description of
- Category nodes: categories reachable in knowledge graph from those topics
- Edge weight = inverse of shortest path length



	c1	*c2*	*c3*	*c4*
t1	$\frac{1}{2}$	1	$\frac{1}{3}$	0
t2	$\frac{1}{2}$	1	$\frac{1}{2}$	1
t3	0	0	$\frac{1}{2}$	1

- Same representation for candidates

Instantiating the Model

- Wikipedia
- DBPedia
- Freebase

Instantiating the Model

- Wikipedia
- DBPedia
- Freebase

Instantiating on Wikipedia

- Articles, categories effectively represent topics (Syed, 2008)

Instantiating on Wikipedia

- Articles, categories effectively represent topics (Syed, 2008)
- Good coverage of even rare entity concepts (Zesch, 2007)

Instantiating on Wikipedia

- Articles, categories effectively represent topics (Syed, 2008)
- Good coverage of even rare entity concepts (Zesch, 2007)
- Compatible with NER toolkits
  - DBPedia Spotlight, Wikipedia Miner

Instantiating on Wikipedia

- Articles, categories effectively represent topics (Syed, 2008)
- Good coverage of even rare entity concepts (Zesch, 2007)
- Compatible with NER toolkits
  - DBPedia Spotlight, Wikipedia Miner
- Article editing behavior effective for modeling interests (Cosley, 2007; Lieberman & Lin, 2009; Wattenberg et al., 2007)

Article editing signals topic interest

Editing behaviors indicative of user interest:

Editing Behavior	Intuition
Number of times user edits article	Repeatedly editing an article implies greater commitment and interest
Article's overall edit activity and total number of editors	Generally popular and actively edited articles are less discriminative of individual interest and personal relevance
Time period user edits article	Long-term interests are stronger than fleeting, short-term interests
Type of edit according to revision tag	Trivial edits such as vandalism reversion or typo correction less indicative of interest than thoughtful, effortful edits
Complexity, completeness, informativeness of edit according to metrics of Information Quality	Type, substantiveness, and overall quality of care user gives to an edit indicates concern and interest in topic

Article editing signals topic interest

Editing behaviors indicative of user interest:

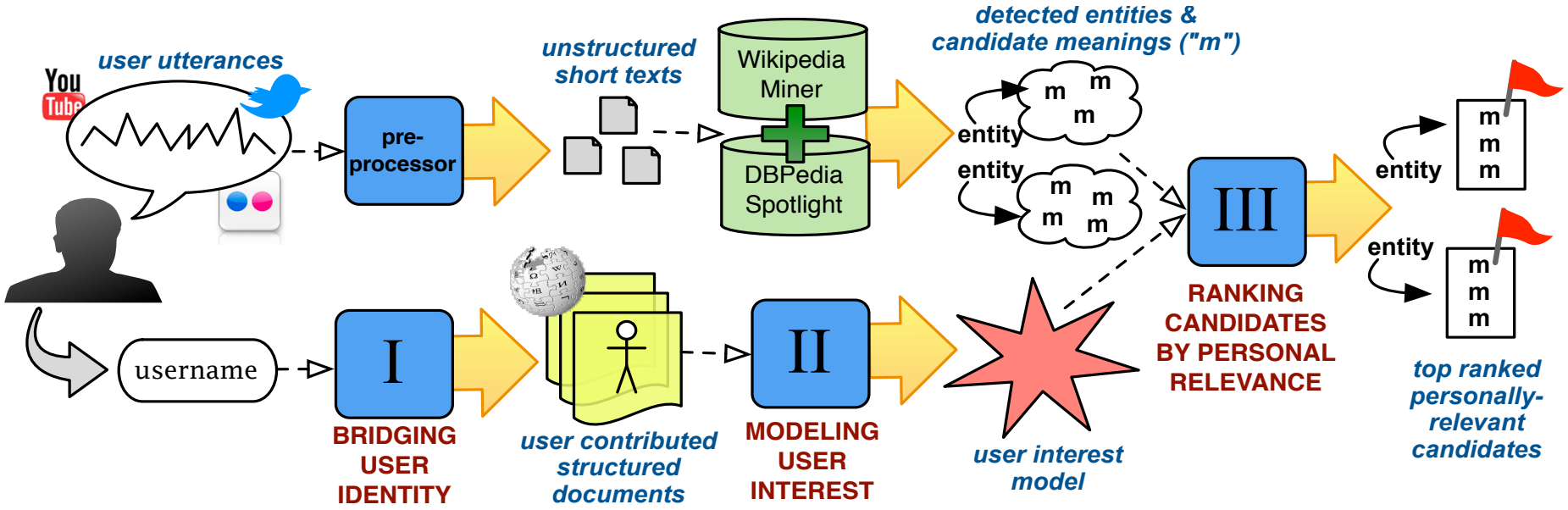
Editing Behavior	Intuition
Number of times user edits article	Repeatedly editing an article implies greater commitment and interest
Article's overall edit activity and total number of editors	Generally popular and actively edited articles are less discriminative of individual interest and personal relevance
Time period user edits article	Long-term interests are stronger than fleeting, short-term interests
Type of edit according to revision tag	Trivial edits such as vandalism reversion or typo correction less indicative of interest than thoughtful, effortful edits
Complexity, completeness, informativeness of edit according to metrics of Information Quality	Type, substantiveness, and overall quality of care user gives to an edit indicates concern and interest in topic

Less Meaningful Edits

Ignore Irrelevant Edits	Clean Article Text
Articles with less than 100 non-stopwords	Stem, tokenize, lowercase; remove stopwords, punctuation, non-printable characters.
Trivial edits, i.e., typo correction, vandalism reversion.	Parse Wiki Markup to remove article maintenance information
List pages merely containing widely diverse sets of topics that are all not necessarily indicative of the piece personally relevant to the user	

Implementation: The RESLVE System

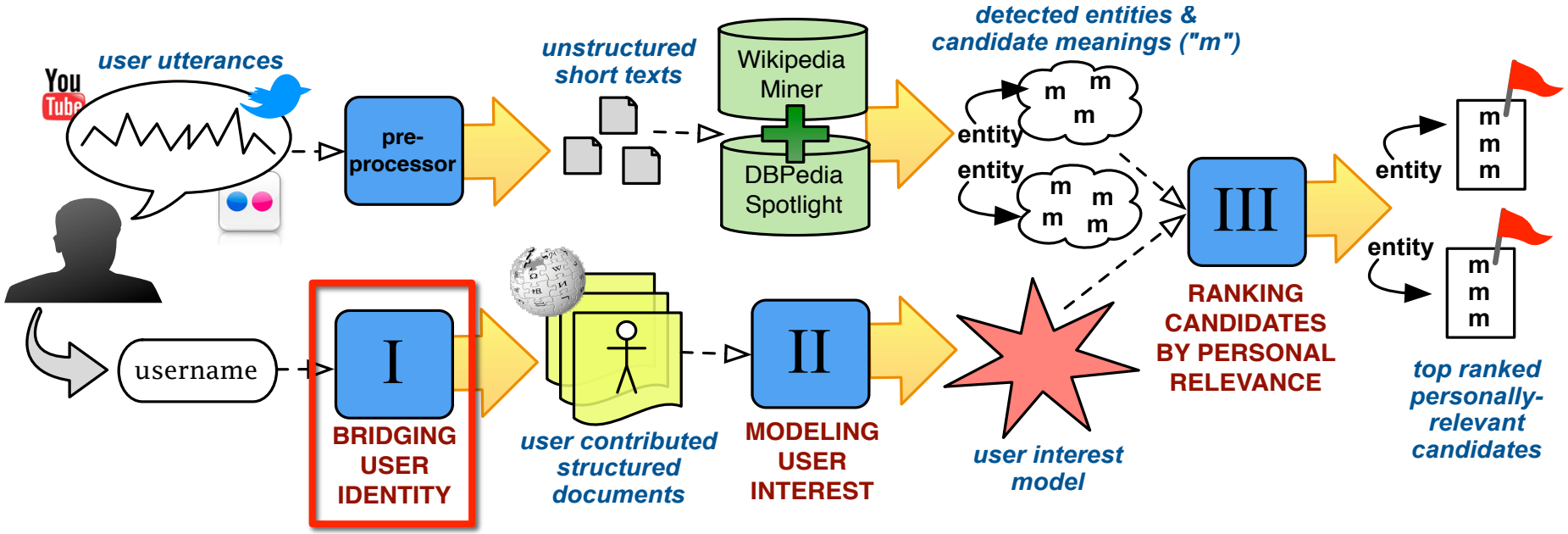
RESLVE (Resolving Entity Sense by LeVeraging Edits) addresses NED by:



Implementation: The RESLVE System

RESLVE (Resolving Entity Sense by LeVeraging Edits) addresses NED by:

1. Connecting social Web + Wikipedia editor identity

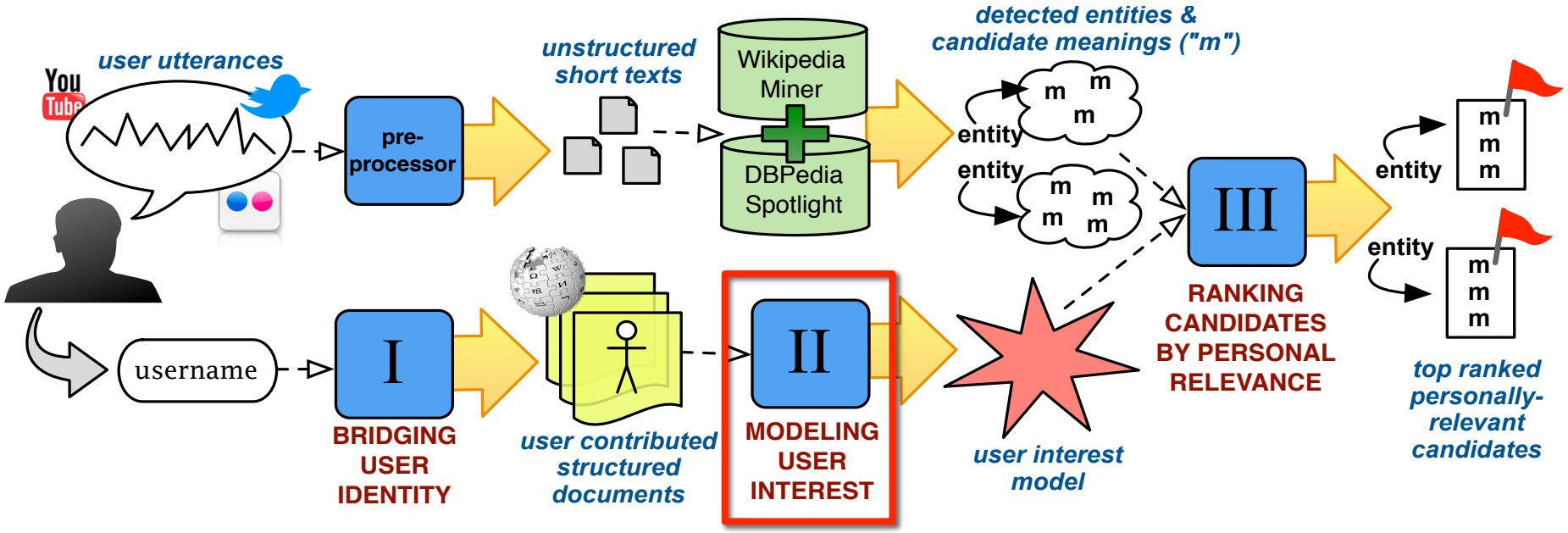


Implementation: The RESLVE System

RESLVE (Resolving Entity Sense by LeVeraging Edits) addresses NED by:

I. Connecting social Web + Wikipedia editor identity

II. Modeling topics of interests using article edits



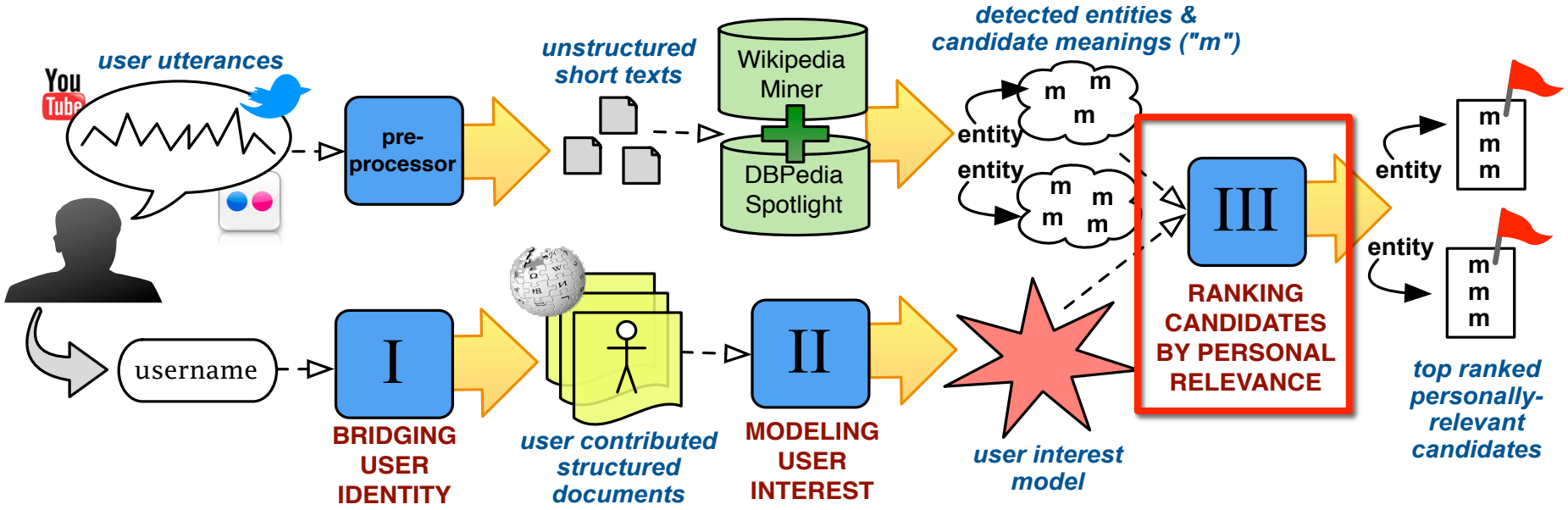
Implementation: The RESLVE System

RESLVE (Resolving Entity Sense by LeVeraging Edits) addresses NED by:

I. Connecting social Web + Wikipedia editor identity

II. Modeling topics of interests using article edits

III. Ranking entity candidates by personal relevance



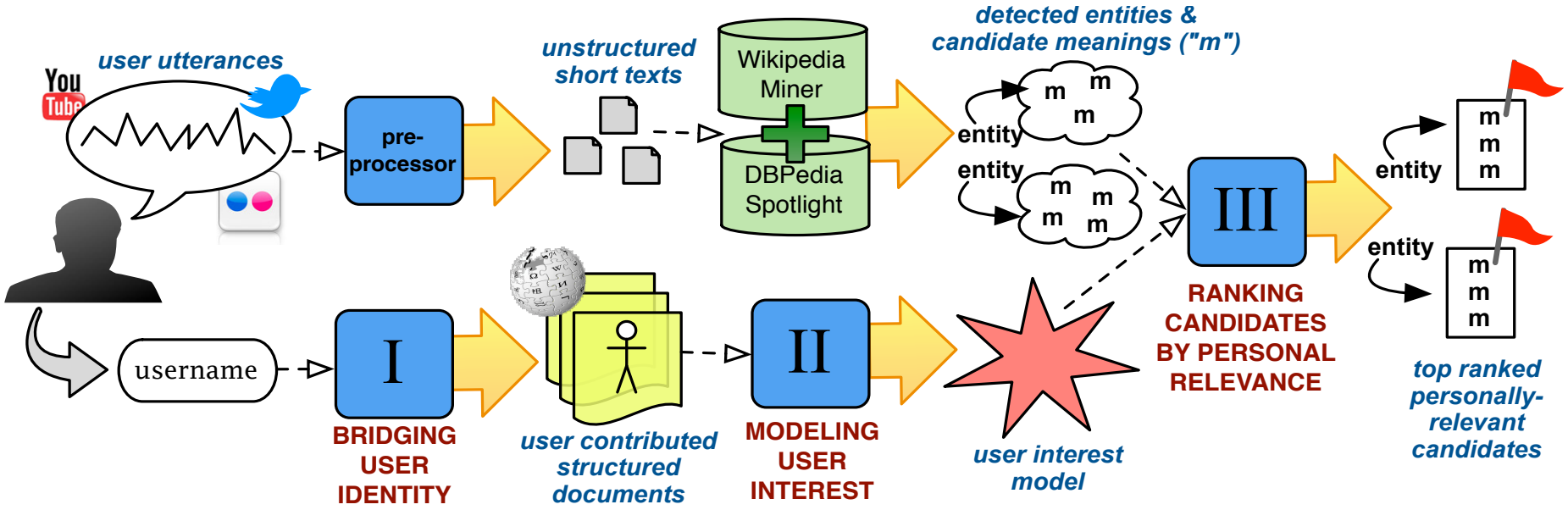
Implementation: The RESLVE System

RESLVE (Resolving Entity Sense by LeVeraging Edits) addresses NED by:

I. Connecting social Web + Wikipedia editor identity

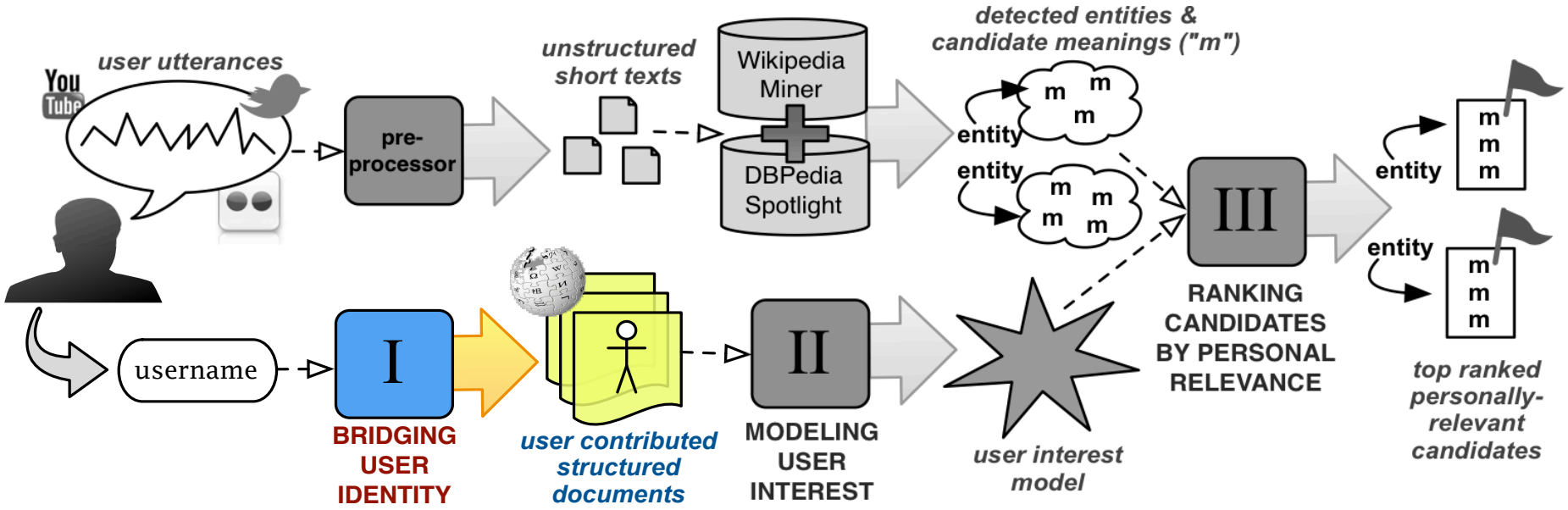
II. Modeling topics of interests using article edits

III. Ranking entity candidates by personal relevance



Phase 1: Bridging Web Identities

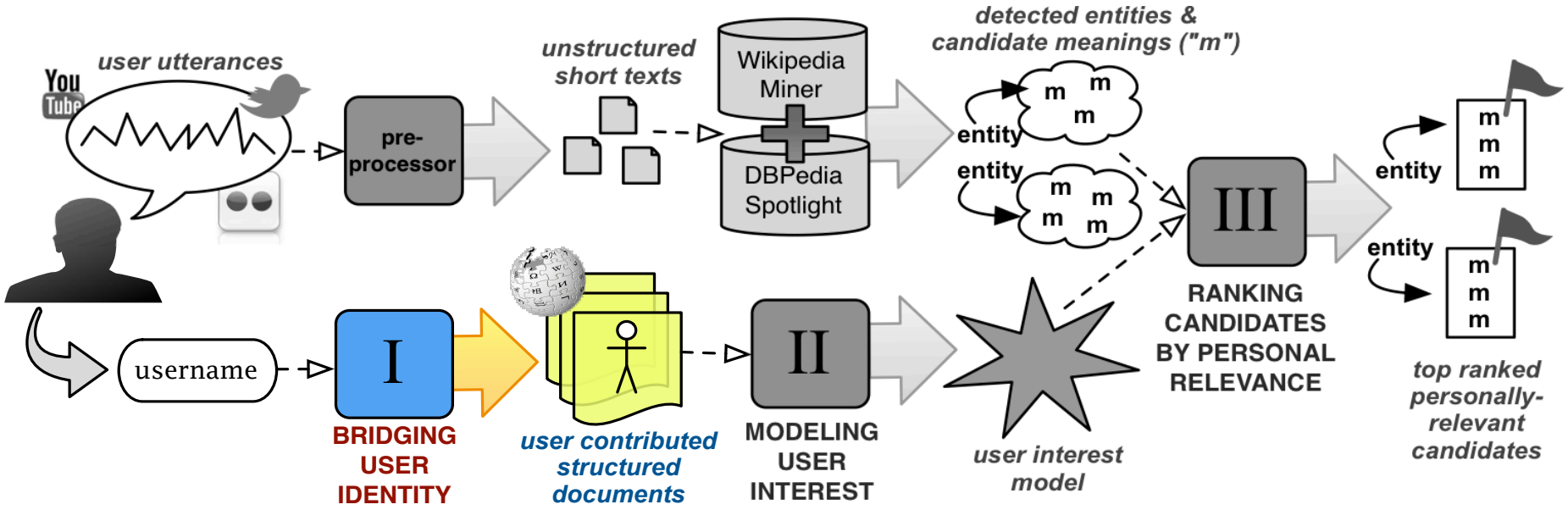
- Connect identity of social media user with Wikipedia editor



Phase 1: Bridging Web Identities

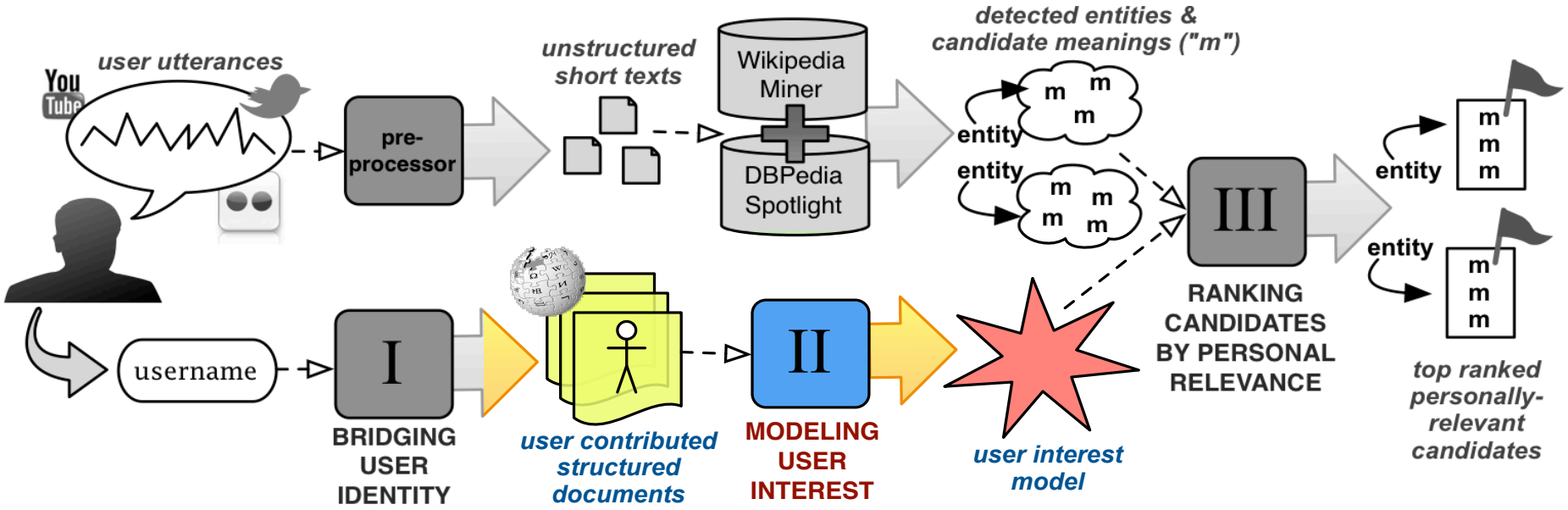
- Connect identity of social media user with Wikipedia editor

- Simple string matching
  - Iofciu, 2011; Perito, 2011



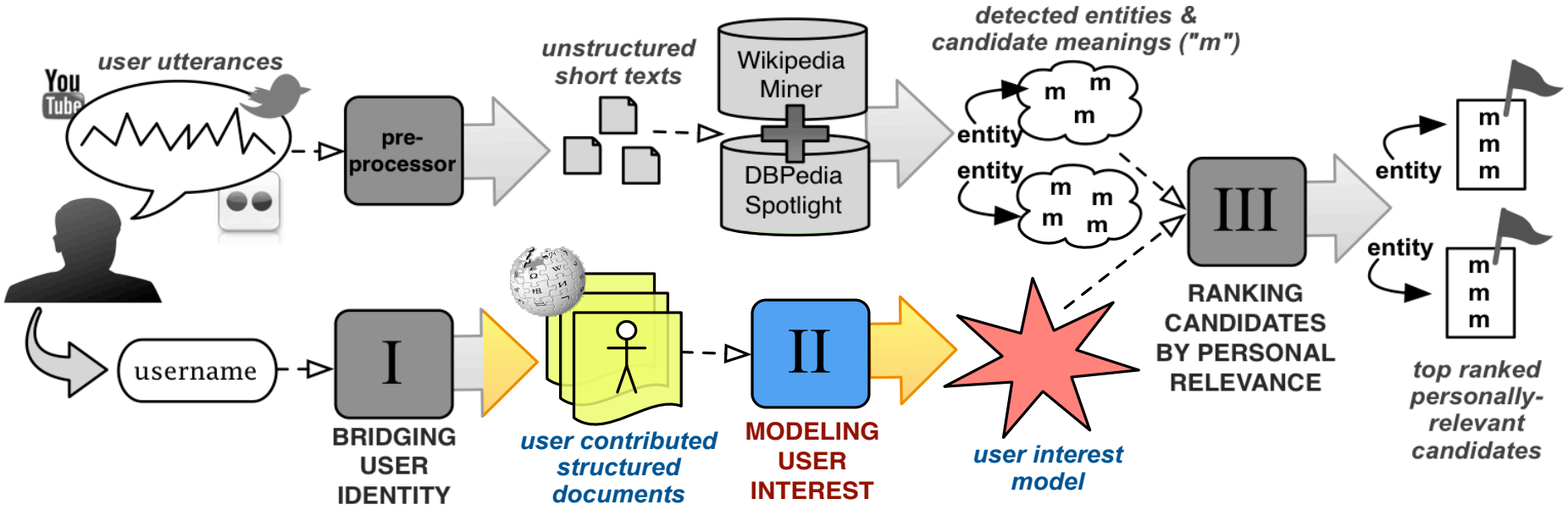
Phase 2: Representing Users and Entities

- Models user's topics of interest using bridged Wiki account's editing-history
- Compares similarity of those topics to topic associated with candidate sense



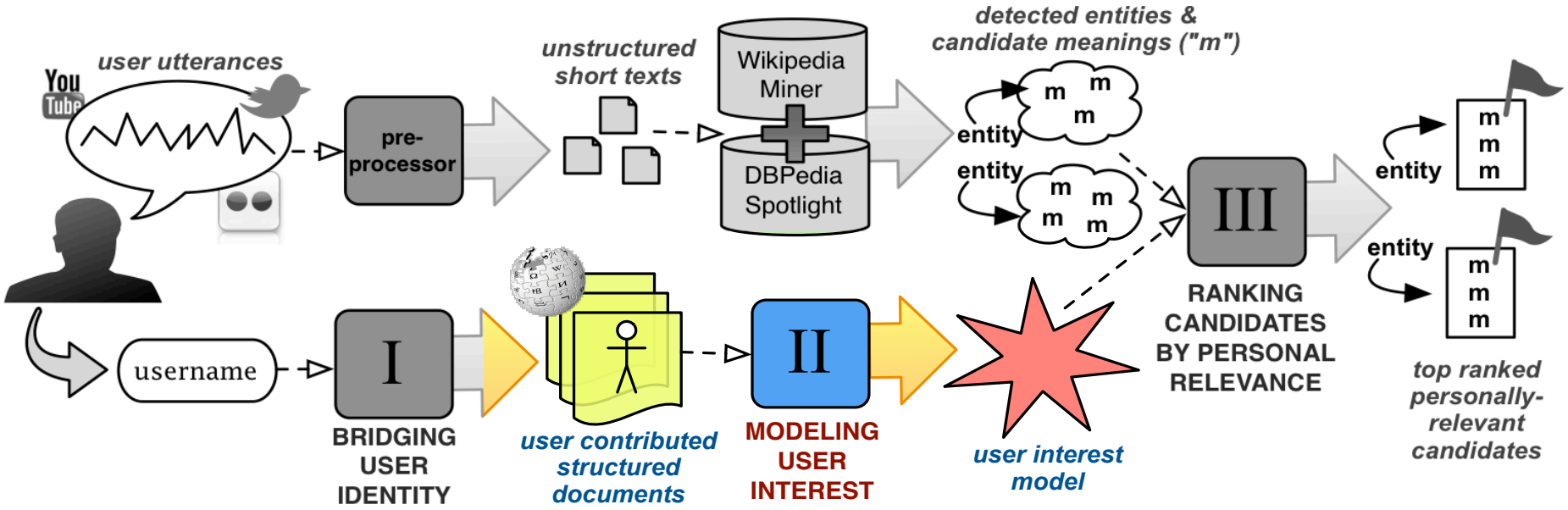
Phase 2: Representing Users and Entities

- Models user's topics of interest using bridged Wiki account's editing-history
- Compares similarity of those topics to topic associated with candidate sense
- Content-based & knowledge-graph based similarity



Phase 2: Representing Users and Entities

- Models user's topics of interest using bridged Wiki account's editing-history
- Compares similarity of those topics to topic associated with candidate sense
- Content-based & knowledge-graph based similarity
- Weighted vectors used to represent user and candidate sense



Content-based similarity

- Bag-Of-Words
  - Titles of articles user edited
  - Candidate's article title
  - Words from those articles' pages & category titles
- TF-IDF weighted

Content-based similarity

- Bag-Of-Words
  - Titles of articles user edited
  - Candidate's article title
  - Words from those articles' pages & category titles
- TF-IDF weighted

- User, u: $V_{content,\ u}$
- Candidate meaning, m: $V_{content,\ m}$

$$sim_{content}(u, m) = cossim(V_{content,\ u}, V_{content,\ m})$$

Knowledge-context based similarity

- Vectors of articles' category IDs
- Weight is distance between the article (topic) and category in knowledge graph
- E.g., "American Television Series" > "Broadcasting"

Knowledge-context based similarity

- Vectors of articles' category IDs
- Weight is distance between the article (topic) and category in knowledge graph
- E.g., "American Television Series" > "Broadcasting"

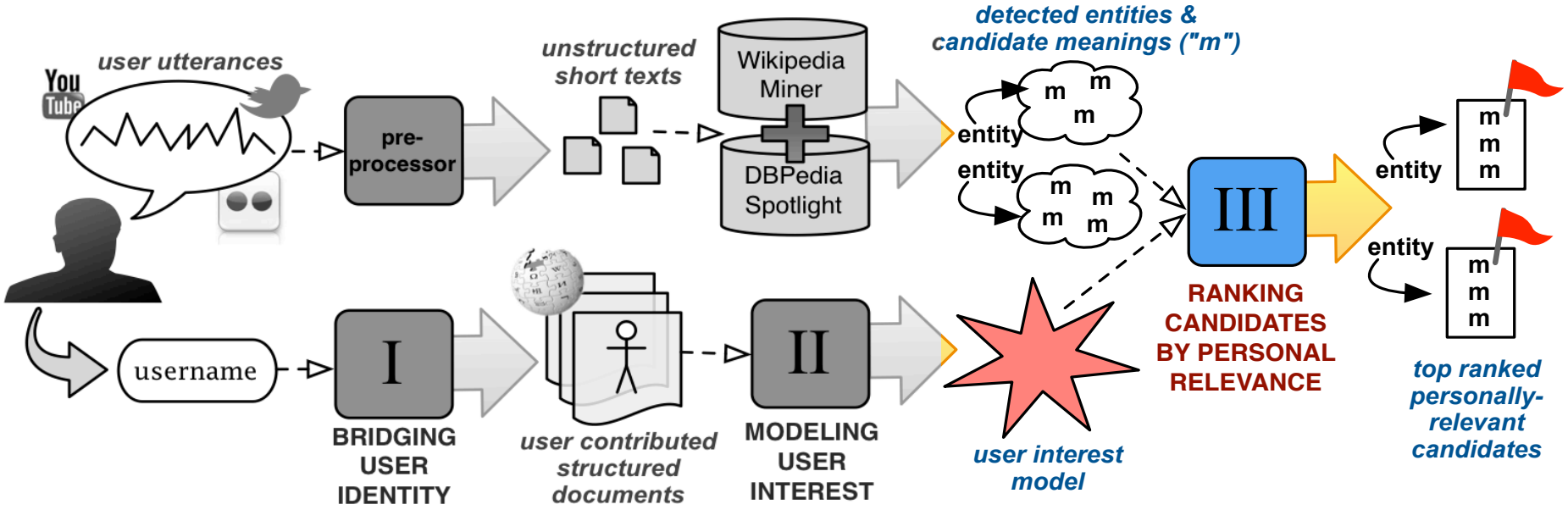
- User, u : $V_{category,\ u}$
- Candidate meaning, m: $V_{category,\ m}$

$$sim_{category}(u, m) = cossim(V_{category,\ u}, V_{category,\ m})$$

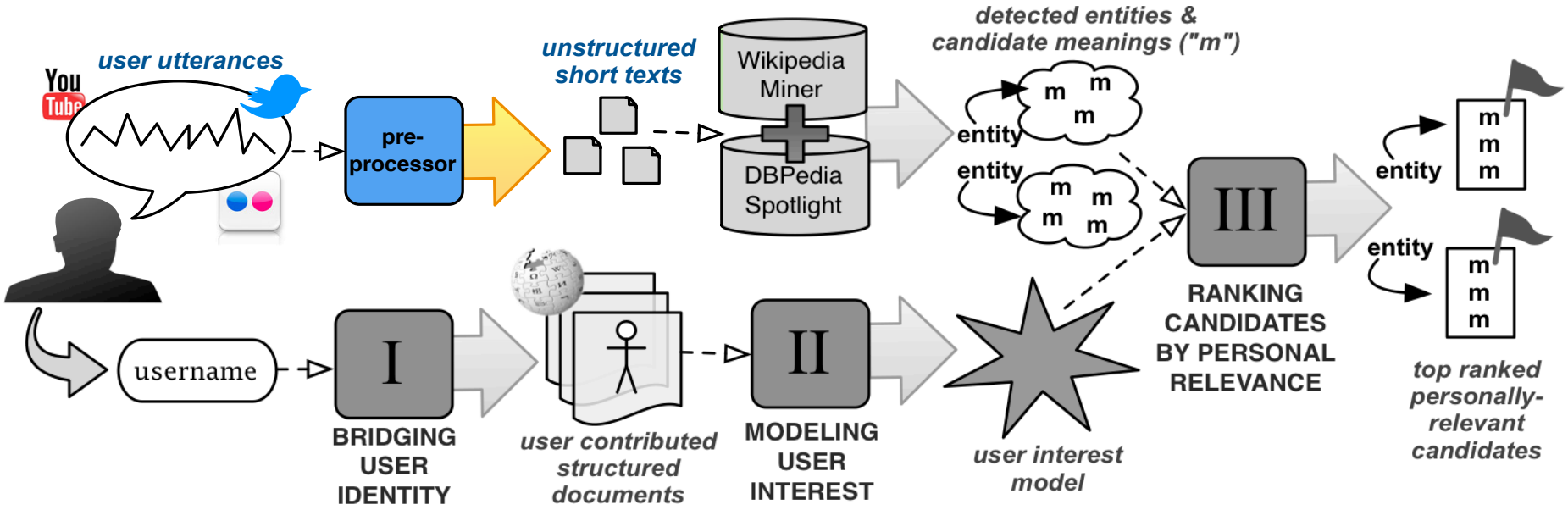
Phase 3: Ranking by Personal Relevance

Output highest scoring candidate as intended meaning by measuring:

$$sim(u,m)=\alpha*sim_{content}(u,m)+(1-\alpha)*sim_{category}(u,m)$$

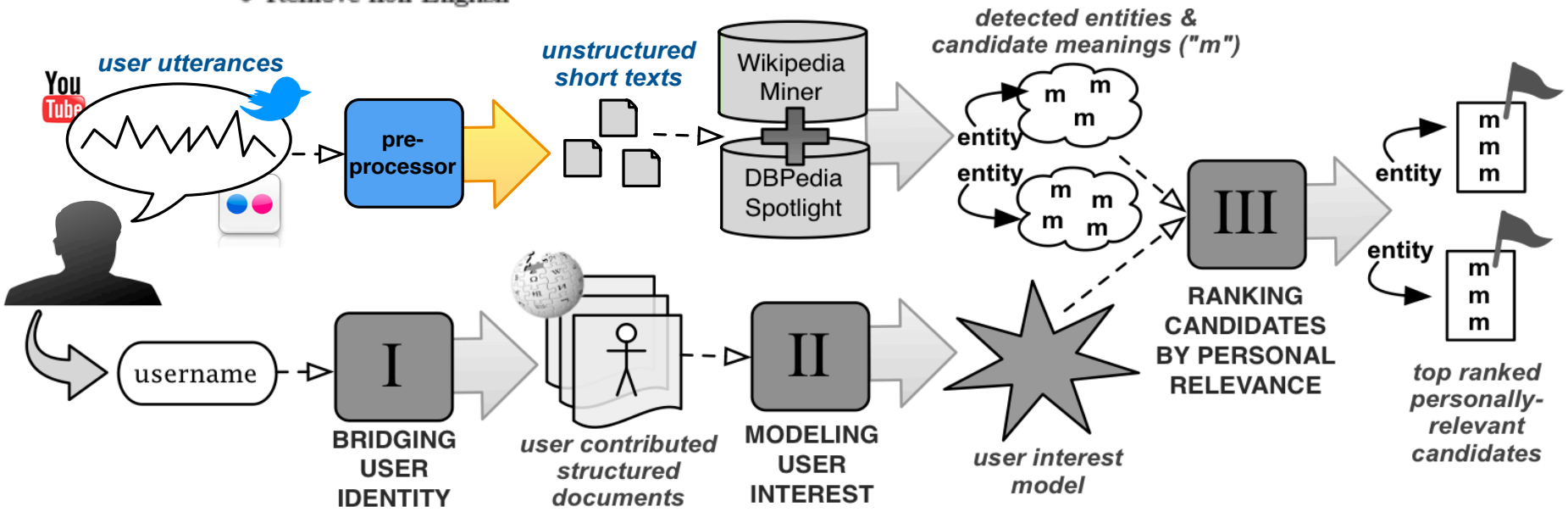


Pre-processing & preparation modules



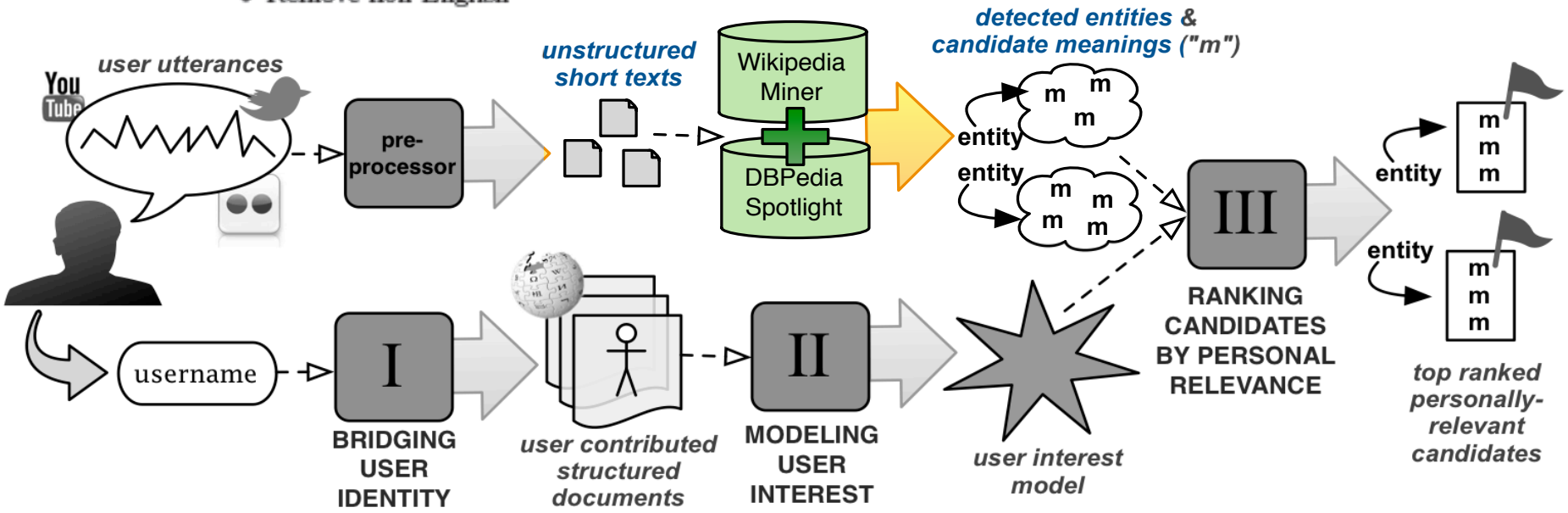
Pre-processing & preparation modules

Tweets:	• Normalize @name to MENTION • Remove RT (retweet) tag • Remove leading "#" but keep hash tag's target concept if English word
YouTube, Flickr:	• Bypass auto-generated file names like IMG_336.jpg or MOV_02.AVI • Remove file type suffix, e.g. ".png", but leave file name if an English word • Ignore auto-generated tags, e.g., "hidden:filter=Boost" machine-tag on Flickr
All utterances:	• Remove URLs • Remove non-English



Pre-processing & preparation modules

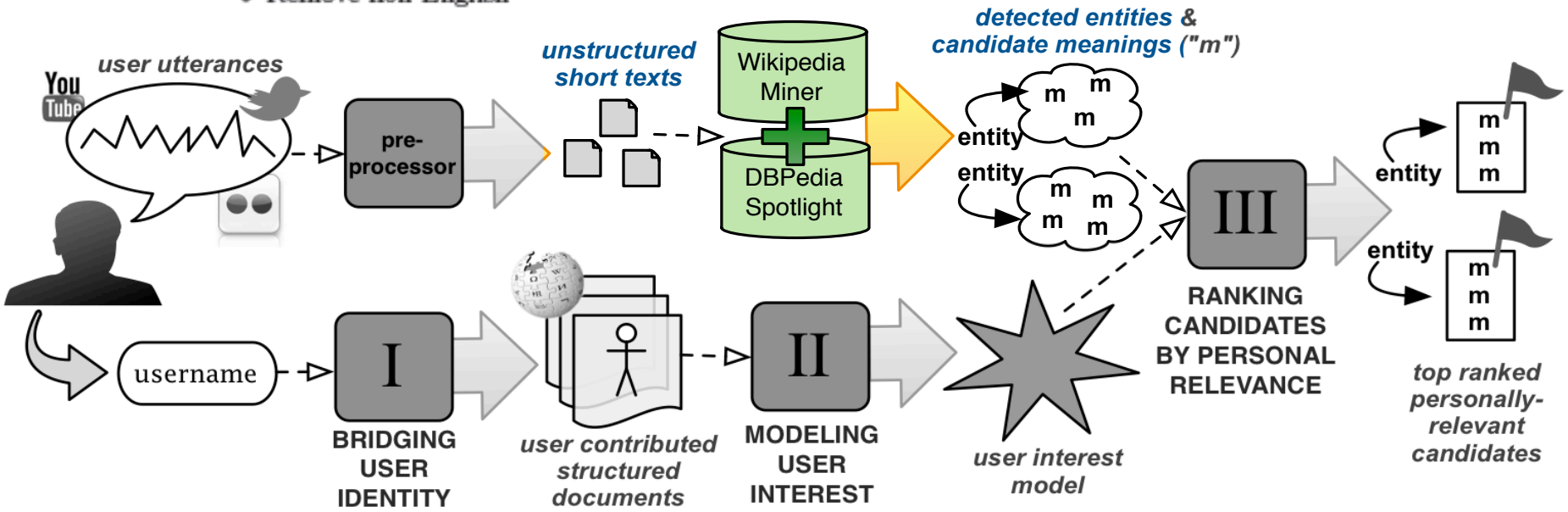
Tweets:	• Normalize @name to MENTION • Remove RT (retweet) tag • Remove leading "#" but keep hash tag's target concept if English word
YouTube, Flickr:	• Bypass auto-generated file names like IMG_336.jpg or MOV_02.AVI • Remove file type suffix, e.g. ".png", but leave file name if an English word • Ignore auto-generated tags, e.g., "hidden:filter=Boost" machine-tag on Flickr
All utterances:	• Remove URLs • Remove non-English



Pre-processing & preparation modules

Tweets:	• Normalize @name to MENTION • Remove RT (retweet) tag • Remove leading "#" but keep hash tag's target concept if English word
YouTube, Flickr:	• Bypass auto-generated file names like IMG_336.jpg or MOV_02.AVI • Remove file type suffix, e.g. ".png", but leave file name if an English word • Ignore auto-generated tags, e.g., "hidden:filter=Boost" machine-tag on Flickr
All utterances:	• Remove URLs • Remove non-English

Language based:	• Non-English • Single characters and parse errors
Entity based:	• Non-entities, i.e., detected terms that are not a Noun class (NN, NNS, NNP, NP) or Named Entity class (e.g., location, person, organization) according to named entity corpora IEER, ACE, or CoNLL • Non-ambiguous entities (0 or 1 meaning)



Agenda

Background

- Task Definitions
- Challenges & Examples
- Attempted Solutions

Approach

- Motivations
- Modeling a Knowledge Context
- Implementation: The RESLVE System

Evaluation

- Experiments
- Results
- Future Work

Experiment

Data Sample

- Twitter: tweets
- YouTube: video titles, descriptions
- Flickr: photo tags, titles, descriptions

Experiment

Data Sample

- Twitter: tweets
- YouTube: video titles, descriptions
- Flickr: photo tags, titles, descriptions

- String-matched usernames of posters to Wikipedia accounts
- Mechanical Turk used to confirm accounts were same person

Experiment

Data Sample

- Twitter: tweets
- YouTube: video titles, descriptions
- Flickr: photo tags, titles, descriptions

- String-matched usernames of posters to Wikipedia accounts
- Mechanical Turk used to confirm accounts were same person

For confirmed matches:

- Collected 100 most recent utterances
- ID, title, page content, categories of edited articles

Experiment

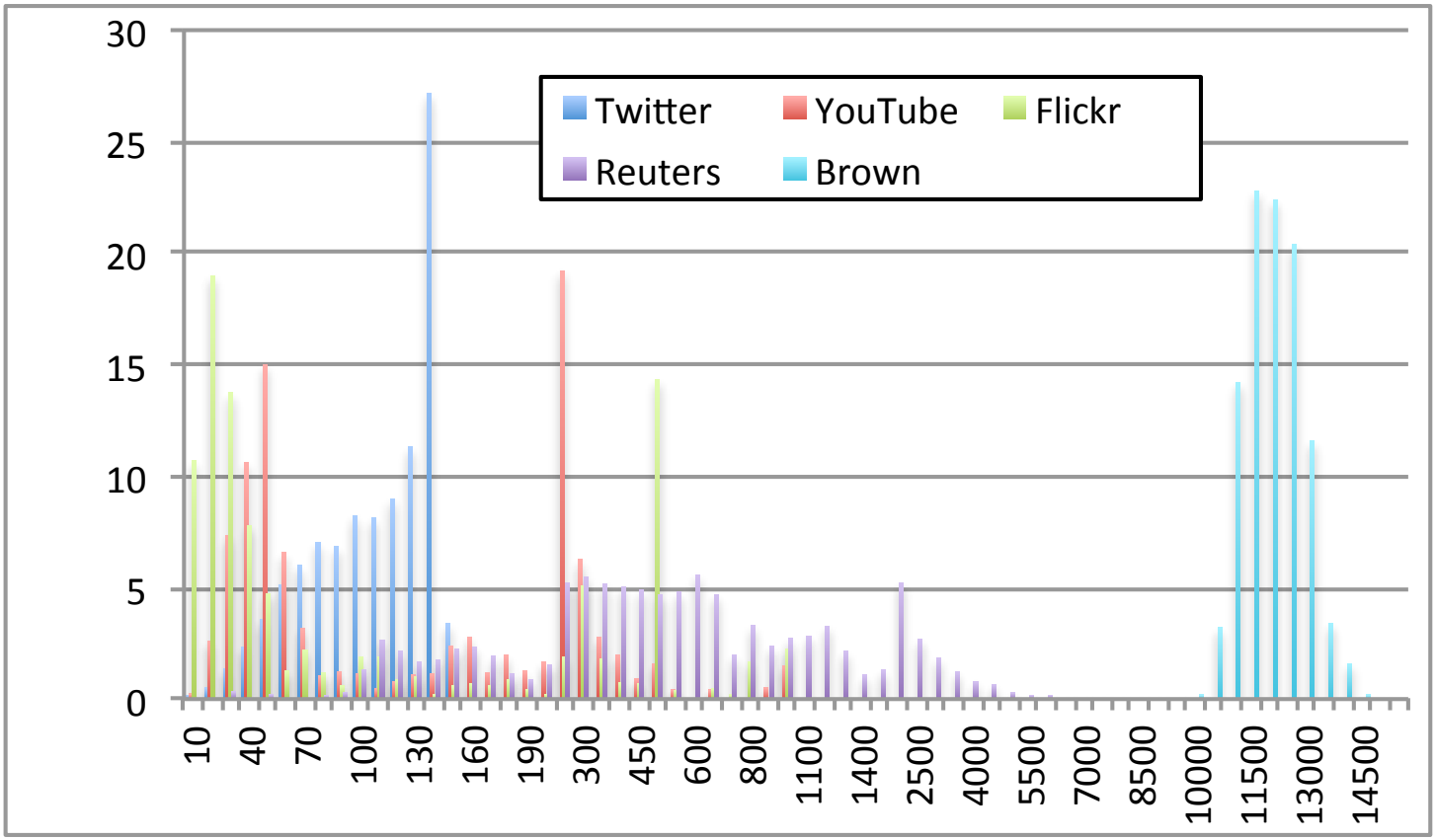
<u>Labeling correct entity meaning</u>

- 1545 valid ambiguous entities
- Mechanical Turk Categorization Masters
- Averaged observed agreement across all coders and items = 0.866
- Average Fleiss Kappa = 0.803
- 918 unanimously labeled ambiguous entities

Dataset Characteristics

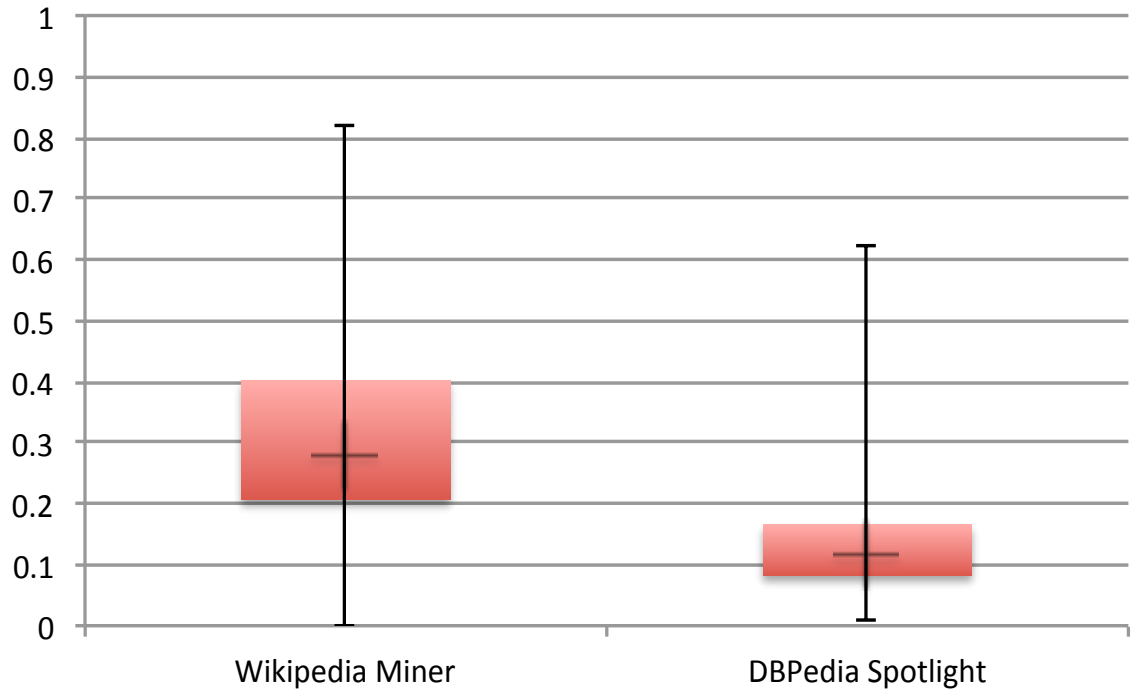
Text Length

Longest utterances still shorter than even shortest texts from NER task corpora like Reuters-21578, Brown-Corpus



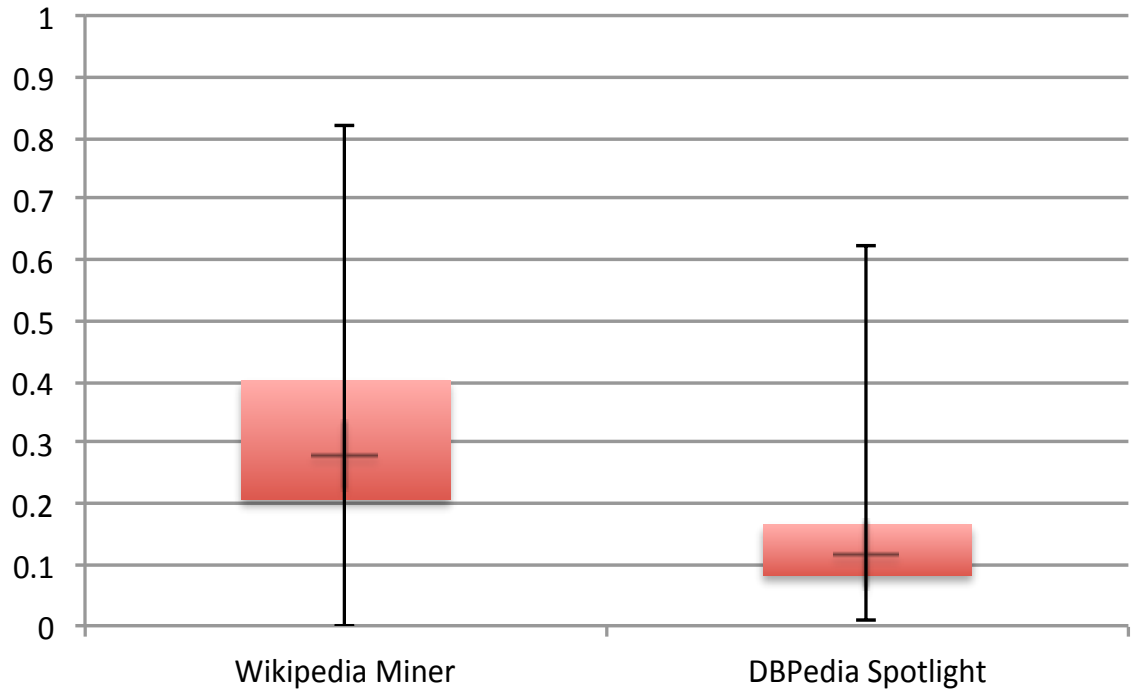
High Ambiguity

- NER services have low confidence



High Ambiguity

- NER services have low confidence



- Many potential candidates (2 to 163, avg. 5-6, median 4)

High Ambiguity

- 91% of utterances contain at least 1 ambiguous entity
- 2/3 of entities detected are ambiguous
- Almost no entities without at least 2 senses to disambiguate

	Twitter	YouTube		Flickr		
	Tweet	*Title*	*Desc*	*Title*	*Desc*	*Tag*
(a)	93%	88%	98%	92%	97%	77%
(b)	64%	55%	46%	66%	44%	73%

Performance

<u>Metric</u>

- Precision at rank 1 (P@1)

Performance

Metric

- Precision at rank 1 (P@1)

Methods of comparison

- Human annotated gold standard
- RC: Randomly sorted candidates
- PF: Prior frequency
- RU: RESLVE given a random Wikipedia user's interest model
- DS: DBPedia Spotlight
- WM: Wikipedia Miner

Results

	flickr	Twitter	YouTube
RESLVE	0.63	**0.76** ★	**0.84** ★
RC	0.21	0.32	0.31
PF	0.74	0.69	0.66
RU	0.51	0.71	0.78
WM	**0.78** ✖	0.58	0.80
DS	0.53	0.67	0.63

Discussion

- Best performance on YouTube texts (longest) due to content-based sim

Discussion

- Best performance on YouTube texts (longest) due to content-based sim
- Outperforms on more personal text (e.g., tweets)
- Random user model less effective

Discussion

- Best performance on YouTube texts (longest) due to content-based sim
- Outperforms on more personal text (e.g., tweets)
- Random user model less effective
- Less effective on impersonal text (e.g., photo geo-tags)
  - High prior frequency so standard methods suffice
  - Personally-unfamiliar topics so not likely to make Wiki edits about them
  - Stable interests assumption breaks down here

Error Cases

- Automated messages
  - *"I uploaded a **video** on @youtube"* → 1945 European Films

Error Cases

- Automated messages
  - *"I uploaded a **video** on @youtube"* → 1945 European Films
- Entities not in knowledge base
  - *"**Peter** on the dock"*

Error Cases

- Automated messages
  - *"I uploaded a **video** on @youtube"* → 1945 European Films
- Entities not in knowledge base
  - *"**Peter** on the dock"*
- Less prolific contributors

Future Work

Future Work

- Computability
  - Wikipedia has 5M articles, 700K categories → Vector pruning

Future Work

- Computability
  - Wikipedia has 5M articles, 700K categories → Vector pruning
- User identity & modeling interests

Bridging User Accounts

	# Usernames	Exist on Wikipedia
Twitter	479	46.1%
YouTube	454	19.6%
Flickr	226	21.7%

Bridging User Accounts

	# Usernames	Exist on Wikipedia	Matches are same person
Twitter	479	46.1%	47%
YouTube	454	19.6%	48%
Flickr	226	21.7%	71%

Bridging User Accounts

Bridging User Accounts

a. True negative (no identity in knowledge base)

Bridging User Accounts

a. True negative (no identity in knowledge base)

b. False negative (same person, different usernames)

Bridging User Accounts

a. True negative (no identity in knowledge base)

b. False negative (same person, different usernames)

c. False positives (string match, but different people)

Bridging User Accounts

a. True negative (no identity in knowledge base)

- ✓ Collaborative filtering techniques to approximate user's own interests with contributions of social connections

b. False negative (same person, different usernames)

c. False positives (string match, but different people)

Bridging User Accounts

a. True negative (no identity in knowledge base)

- ✓ Collaborative filtering techniques to approximate user's own interests with contributions of social connections

b. False negative (same person, different usernames)

- ✓ Consider more profile attributes than username

c. False positives (string match, but different people)

Bridging User Accounts

a. True negative (no identity in knowledge base)

- ✓ Collaborative filtering techniques to approximate user's own interests with contributions of social connections

b. False negative (same person, different usernames)

- ✓ Consider more profile attributes than username

c. False positives (string match, but different people)

Bridging User Accounts

a. True negative (no identity in knowledge base)

✓ Collaborative filtering techniques to approximate user's own interests with contributions of social connections

b. False negative (same person, different usernames)

✓ Consider more profile attributes than username

c. False positives (string match, but different people)

- Use other knowledge base besides Wikipedia

Bridging User Accounts

a. True negative (no identity in knowledge base)

✓ Collaborative filtering techniques to approximate user's own interests with contributions of social connections

b. False negative (same person, different usernames)

✓ Consider more profile attributes than username

c. False positives (string match, but different people)

- Use other knowledge base besides Wikipedia
- Model user interest from additional kinds of participation (e.g., page visits, bookmarking favoriting)

Bridging User Accounts

a. True negative (no identity in knowledge base)

✓ Collaborative filtering techniques to approximate user's own interests with contributions of social connections

b. False negative (same person, different usernames)

✓ Consider more profile attributes than username

c. False positives (string match, but different people)

- Use other knowledge base besides Wikipedia
- Model user interest from additional kinds of participation (e.g., page visits, bookmarking favoriting)
- Interest drift & time-frame of postings

Summary & Conclusion

- Social Web texts: short & highly personal
- User posts about same topics across communities (but not always)
- Models user interest as personal context with respect to a knowledge base's categorical organization scheme
- Ranking technique compares entity's potential meanings to user's interests to determine intended meaning
  - Language and context independent
- Promising performance gains
- Going forward: such a strategy becomes increasingly necessary, feasible, and effective

Thank You!

Elizabeth L. Murnane
elm236@cornell.edu

Bernhard Haslhofer
bernhard.haslhofer@univie.ac.at

Carl Lagoze
clagoze@umich.edu

Acknowledgements

- Claire Cardie, Dan Cosley, Lillian Lee, Sean Allen, Wenceslaus Lee
- National Science Foundation Graduate Research Fellowship under Grant No. DGE 1144153
- Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (PIOF-GA-2009-252206).

- Questions?