



Implicit Entity Linking in Tweets

Sujan Perera, **Pablo N. Mendes**, Adarsh Alex, Amit P. Sheth,
Krishnaprasad Thirunarayan

Kno.e.sis Center, Wright State University
IBM Research, Almaden

Motivation

- Linking explicitly mentioned entities in tweets is well-explored
- Tweets also contain implicit mentions of entities

Explicit
Entity

Implicit
Entity

 “New **Sandra Bullock** astronaut lost in space **movie** looks absolutely terrifying.”

Implicit
Entity

 “ISRO sends probe to Mars for less money than it takes Hollywood **movie** to send a woman to space.”

Motivation

- Sentiment analysis



“New Sandra Bullock astronaut lost in space movie looks absolutely terrifying.”

Gravity - Positive

- Trend detection



“Kinda sad to hear about that South African runner kill his girlfriend”

Oscar Pistorius

- Event monitoring



“Texas Town Pushes for Marijuana Legalization to Combat Cartel Traffic”

Marijuana Legalization in El Paso

Ignoring implicit entity mentions would adversely affect downstream analysis tasks

Implicit Entities

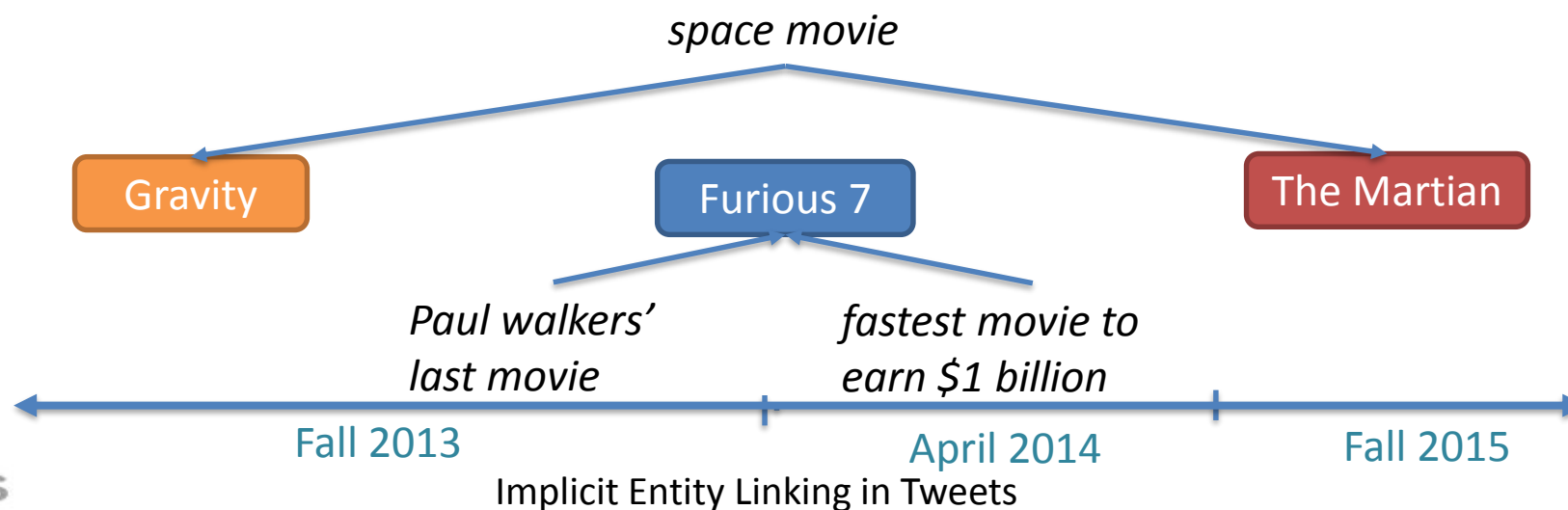
- Definition
 - Implicit entity is an entity mentioned in text where its name is not present nor it is a synonym/alias/abbreviation or a co-reference of an explicitly mentioned entity in the text.
- Prevalence
 - **21%** of movie mentions and **40%** of book mentions are implicit in tweets.
- Implicit entity linking
 - Given a text with an implicit entity mention of a particular type (e.g. Movie, Book, Disorder) output the entity mentioned by the text w.r.t a given knowledge base.

Characteristics

- Types of references-through-characteristics
 - “... Richard Linklater movie ...”
 - “... Ellar Coltrane on his 12-year movie ...”
 - “... 12-year long movie shoot ...”



- Dynamic Context



Implicit Entity Linking in Tweets

Twitter users often rely on sources of context outside the current post, assuming that perhaps there is some shared relationship between them and their audience, or temporal context in the form of recent events and recently mentioned entities (Derczynski et al., 2015)

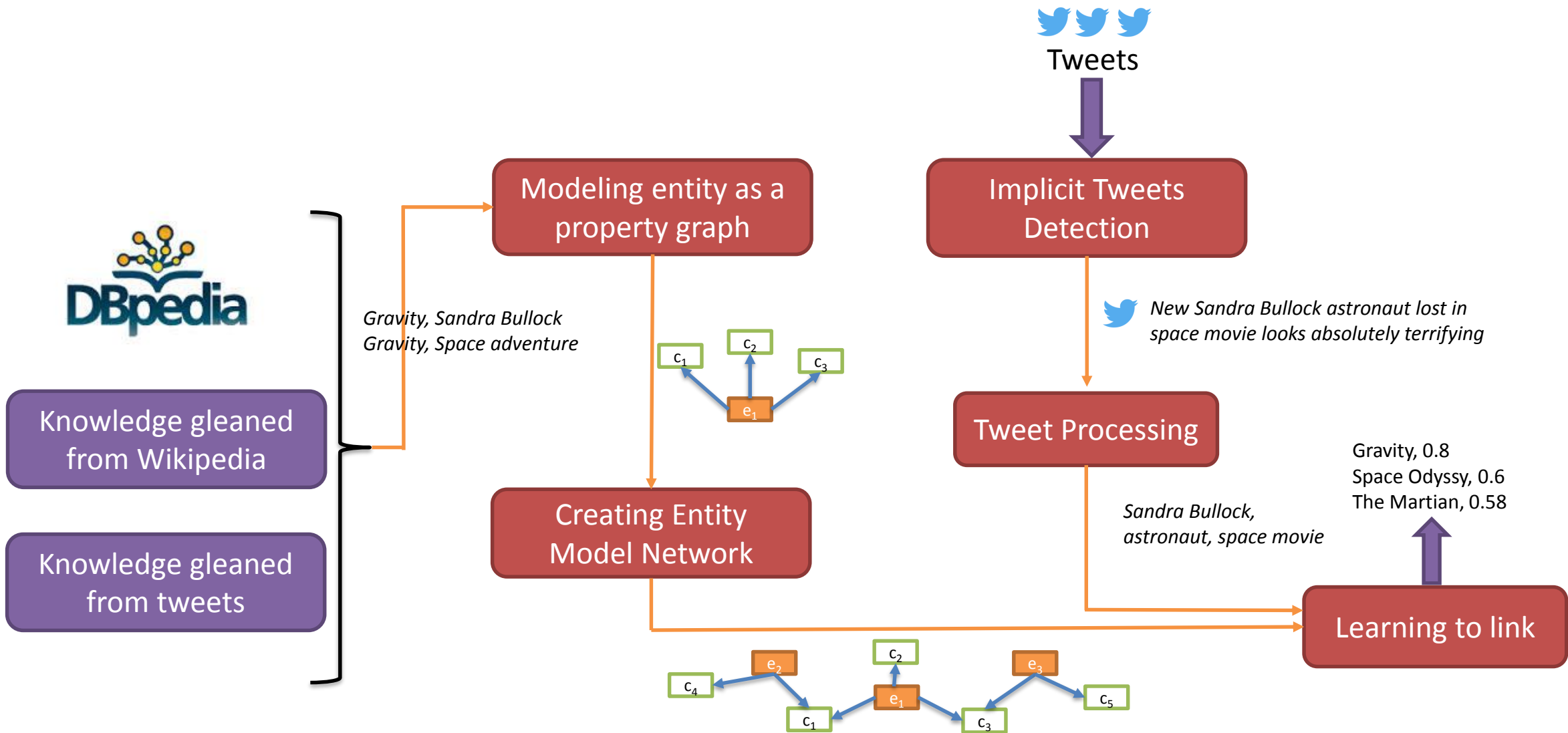
 “New Sandra Bullock astronaut lost in space movie looks absolutely terrifying.” – **Gravity**



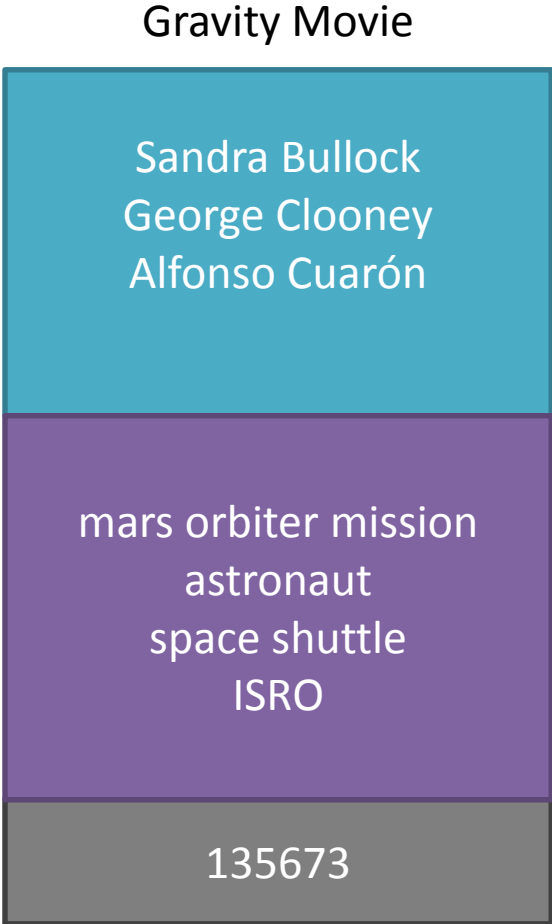
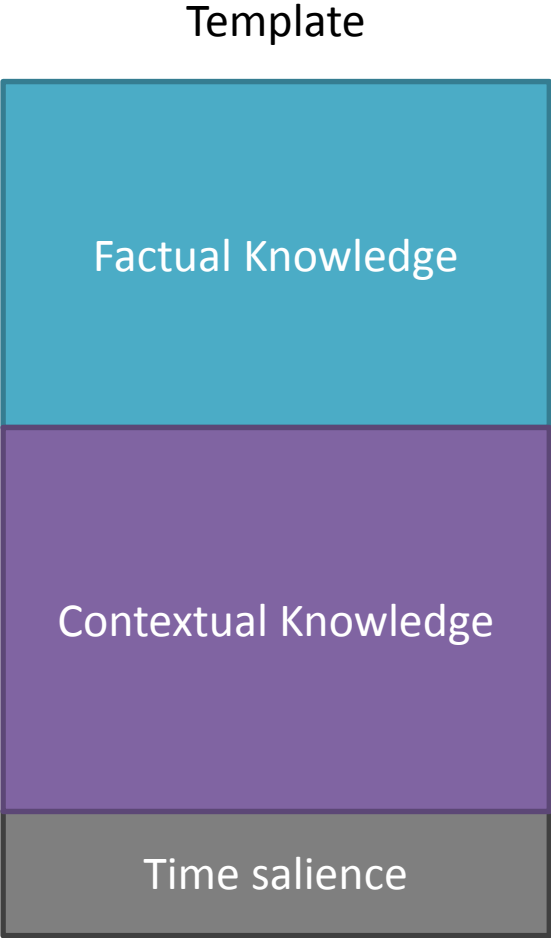
 “ISRO sends probe to Mars for less money than it takes Hollywood to send a woman to space.” - **Gravity**

Indian Space Research Organization’s Mars orbiter mission cost less than the movie Gravity

Implicit Entity Linking in Tweets



Entity Modeling



Knowledge Acquisition

- Acquiring factual knowledge

- Source – DBpedia
- Not all factual knowledge is important – movie has *'starring'* and *'director'* as well as *'billed'* and *'license'*
- Rank the relationships based on joint probability with the entity type

$$P(r, T) = \frac{\text{number of triples of } r \text{ with instances of } T}{\text{total number of triples of } r}$$

Factual
Knowledge

- Acquiring contextual knowledge

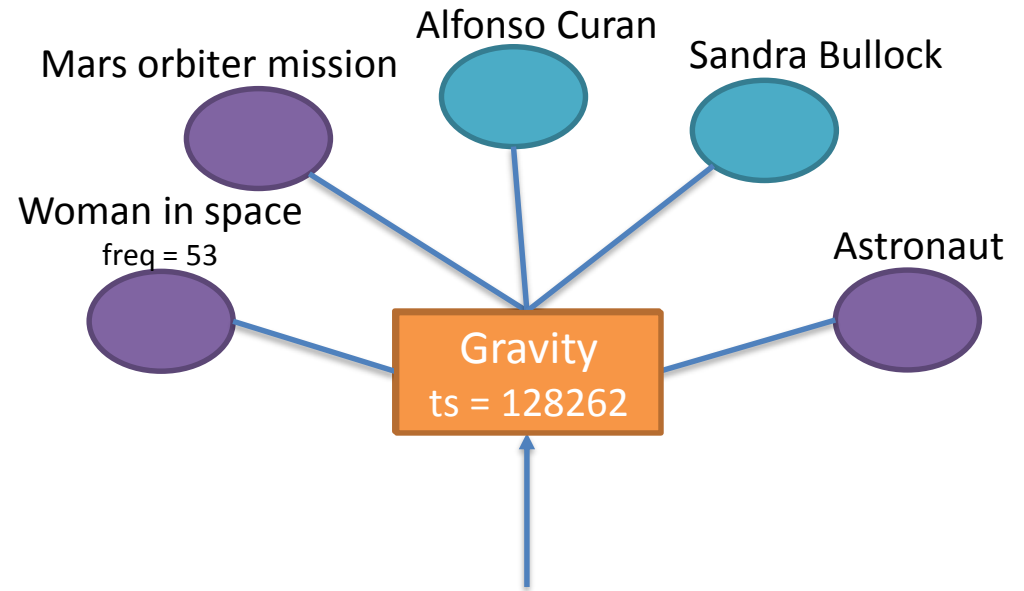
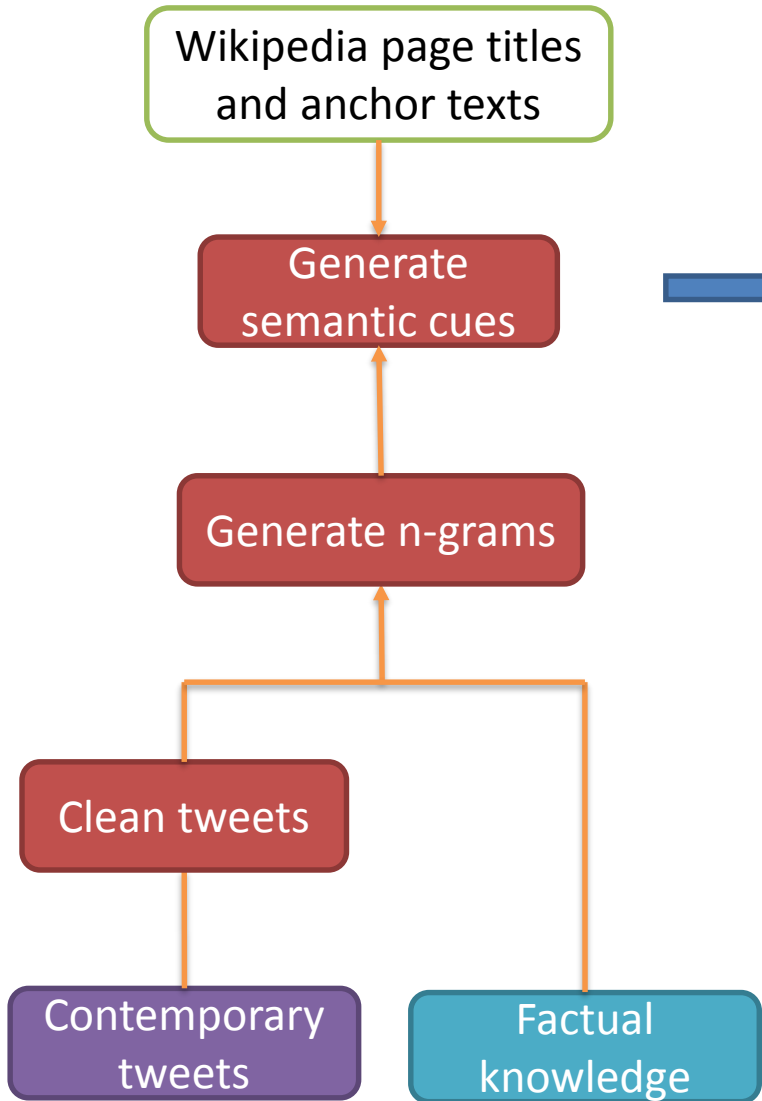
- Source – contemporary tweets
- We collect 1000 tweets with explicit mentions of the entity

Contextual
Knowledge

- Collect the number of hits for the entity's Wikipedia page within last t days as its temporal salience

Time
salience

Entity Modeling



Wikipedia Page Views



WIKIPEDIA
The Free Encyclopedia



Factual Knowledge



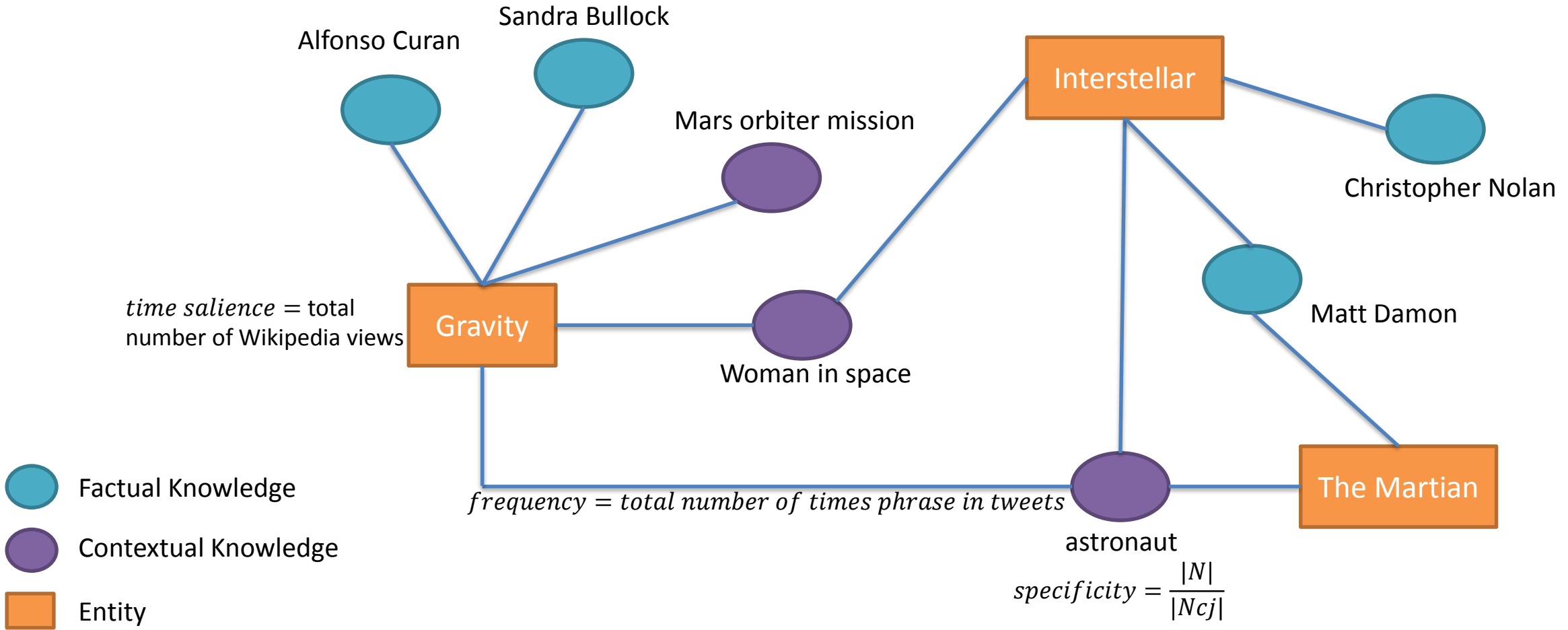
Contextual Knowledge



Entity

Entity Model Network

- A property graph - reflecting the topical relationships between entities



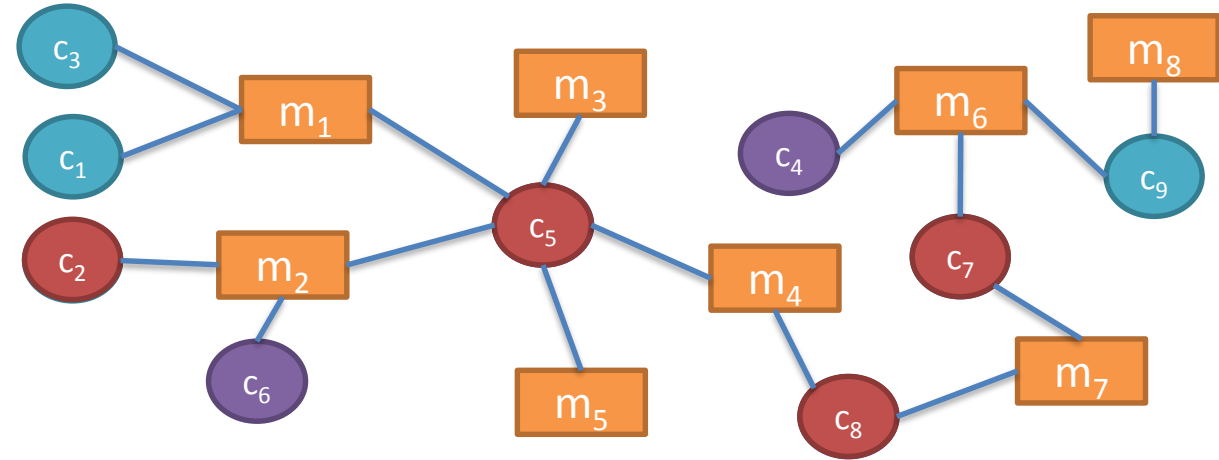
N – total number of entities, *Ncj* number of adjacent entities

Entity Linking

- Two Step Process
- Step 1: Candidate selection and filtering
 - Objective - prune the search space to reduce number of entities should be considered in disambiguation step from EMN
- Step 2: Disambiguation
 - Objective - sort the selected candidate entities to place the implicitly mentioned entity in top position

Learning to Link - Candidate Selection and Filtering

 *ISRO sends probe to Mars for less money than it takes Hollywood to send a woman to space*



- m₁
- m₂
- m₃
- m₄
- m₅
- m₆
- m₇

$$score_{m_i} = \sum_{c_j \in \mathbb{C}} \text{specificity of } c_j * \text{frequency}(c_j, m_i)$$

\mathbb{C} is the set of matching cues

- m₂
- m₄
- m₆
- m₇
- m₃

 Factual Knowledge  Contextual Knowledge  Entity

Learning to Link - Disambiguation

- Formulated as a ranking problem
- SVM^{rank} to rank candidates
 - Similarity between the candidate entity and the tweet

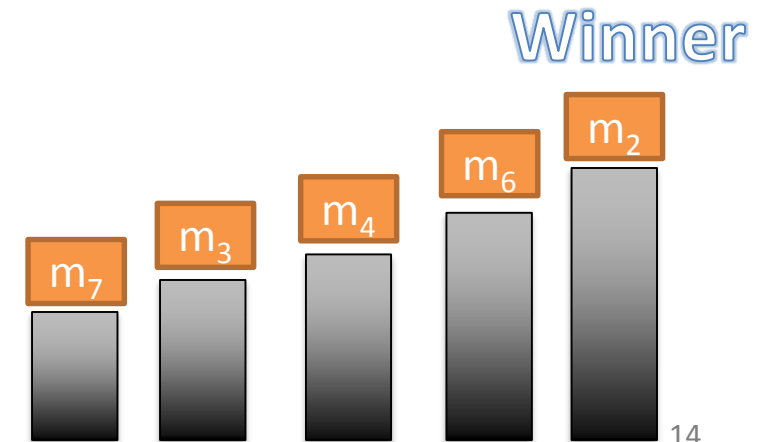


$$x_j = \text{specificity of } c_j * \text{frequency}(c_j, e_i)$$

- Time salience of the candidate entity

$$\frac{\text{temporal salience } e_i}{\sum_{e \in E_c} \text{temporal salience } e}$$

E_c is the selected candidate set



Evaluation Dataset

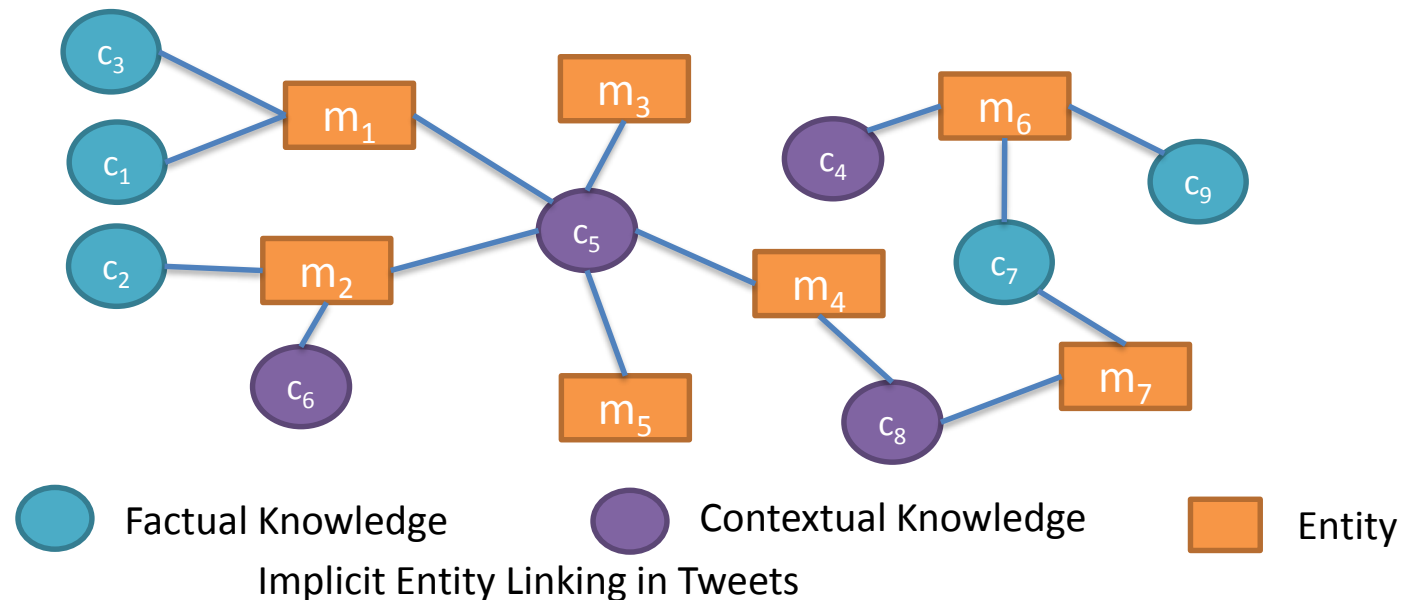
- Manually annotated tweets with two entity types
- Tweets are collected in August 2014 – using keywords ‘movie’ and ‘film’ for movies and ‘book’ and ‘novel’ for books

Entity Type	Annotation	Tweets	Entity
Movie	Explicit	391	107
	Implicit	207	54
	NIL	117	0
Book	Explicit	200	24
	Implicit	190	53
	NIL	70	0

- The tweets annotated with NIL do not have either explicit or implicit mention of an entity

Entity Model Network for Evaluation

- 15,000 tweets for movies and books in July 2014
 - 617 movies and 102 books
- Recent 1000 tweets per entity to build its contextual knowledge
- May 2014 version of DBpedia used to extract factual knowledge
- Temporal salience is obtained for July 2014



Evaluation - Implicit Entity Linking

- How many tweets had correct entity within selected candidate set (top-25) ?
- How many entities were correctly linked by our disambiguation approach?

Entity Type	Candidate Selection Recall	Disambiguation accuracy
Movie	90.33%	60.97%
Book	94.73%	61.05%

- Importance of contextual knowledge

Step	Entity Type	Without ctx	With ctx
Candidate Selection Recall	Movie	77.29%	90.33%
	Book	76.84%	94.73%
Disambiguation Accuracy	Movie	51.7%	60.97%
	Book	50.0%	61.05%

Qualitative Error Analysis

Error	Tweet	Entity
Lack of contextual knowledge	<i>"That Movie Where Shailene Woodley Has Her First Nude Scene? The Trailer Is RIGHT HERE!: No one can say Shailene Woodley isn't brave!"</i>	White Bird in a Blizzard
Novel entities	<i>"hey, what's wrawng widdis goose?" RT @TIME: Mark Wahlberg could be starring in a movie about the BP oil spill http://ti.me/1oZh55V"</i>	Deepwater Horizon
Cold start of entities	<i>"Video: George R.R. Martin's Children's Book Gets Re-release http://bit.ly/1qNNH5r"</i>	The Ice Dragon
Multiple implicit entity mentions	<i>"That moment when you realize that hazel grace and Augustus are brother and sister in one movie and in love battling cancer in another"</i>	Divergent, The Fault in Our Stars

Conclusion and Future Work

- Introduced a novel task and studied its characteristics
- Developed a knowledge-driven solution
- Implement operators to capture evolving knowledge
 - A new entity becomes popular and people start to tweet about it or the popularity of an existing entity fades away
 - A new topic of interest emerges for an existing entity or with the introduction of a new entity, or the popularity of the existing topic fades away
- Develop technique to identify the tweets with implicit entity mentions and their type (“recognition” as in NER)
- Expand the evaluation to other domains and use larger datasets

Thank You

<http://www.knoesis.org>

Dataset is available at:

<https://goo.gl/jrwpeo>