# Leveraging Knowledge Graphs for Web-Scale Unsupervised Semantic Parsing

LARRY HECK, DILEK HAKKANI-TÜR, GOKHAN TUR

# Focus of This Paper

## *SLU and Entity Extraction (Slot Filling)*

**Spoken Language Understanding (SLU):** convert automatic speech recognizer (ASR) output into pre-determined semantic output format

**DOMAIN = movies**

"when was james cameron's avatar released"

INTENT: Find_release_date
MOVIE NAME: avatar
DIRECTOR NAME: james cameron

| Intents | Slots |
|---|---|
| Find movie | Movie genre |
| Find showtime | Movie award |
| Find theater | Theater location |
| Buy tickets | Number of tickets |
| ... | ... |

**DOMAIN = company**

"show me media companies in california"

INTENT: Find_company
LOCATION: california
INDUSTRY: media

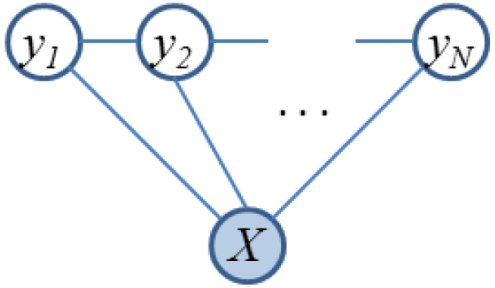
| Intents | Slots |
|---|---|
| Find company | Company name |
| Find revenue | Company address |
| Find founder | Company revenue |
| Find contact | Company industry |
| ... | ... |

# Modeling Entities (Slots) for Semantic Parsing

- Typically framed as a sequence classification problem, where CRFs are shown to be suitable:

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\, P(Y|X)$$

- A linear chain CRF with first order Markov constraint




$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_k \lambda_k f_k(y_{t-1}, y_t, x_t)\right)$$

| | *show* | *me* | *recent* | *action* | *movies* | *by* | *james* | *cameron* |
|---|---|---|---|---|---|---|---|---|
| ***Slots*** | O | O | B-date | B-genre | O | O | B-director | I-director |
| ***Intent*** | Find Movie | | | | | | | |

# Overview

**Problem Statement**

- Developing semantic parsing in SLU typically requires manual crafting for each domain (schema, data, collection, annotations)

- As a result…
  - Narrow breadth of domains
  - Limiting sharing of data/schemas between domains
  - Limited ability to incorporate disparate knowledge sources
  - Inflexible to changes in task definition

How can we reduce the time to create and deploy SLU for a given domain?

# Approach

**Domain independent SLU:** bootstrap semantic parser (slot filling) from web-scale parser

**Knowledge as Priors:** Leverage large knowledge/semantic graphs (e.g., Freebase) to bootstrap web-scale semantic parsers

Use **unsupervised learning** to enable web-scale domain coverage

- No semantic schema design
- No data collection
- No manual annotations


[1] Larry Heck and Dilek Hakkani Tur, Exploiting the Semantic Web for Unsupervised Spoken Language Understanding, IEEE SLT Workshop, December 2012

[2] Gokhan Tur, Minwoo Jeong, Ye-Yi Wang, Dilek Hakkani-Tur, and Larry Heck, Exploiting the Semantic Web for Unsupervised Natural Language Semantic Parsing, in *Proceedings of Interspeech*, International Speech Communication Association, September 2012

[3] Dilek Hakkani-Tur, Larry Heck, and Gokhan Tur, Using a Knowledge Graph and Query Click Logs for Unsupervised Learning of Relation Detection, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013

# Semantic Knowledge Graphs

## *Important Priors for SLU*

- **What are Knowledge Graphs (KGs)?**
  *Graph of strongly typed nodes (entities) connected by edges (properties).*

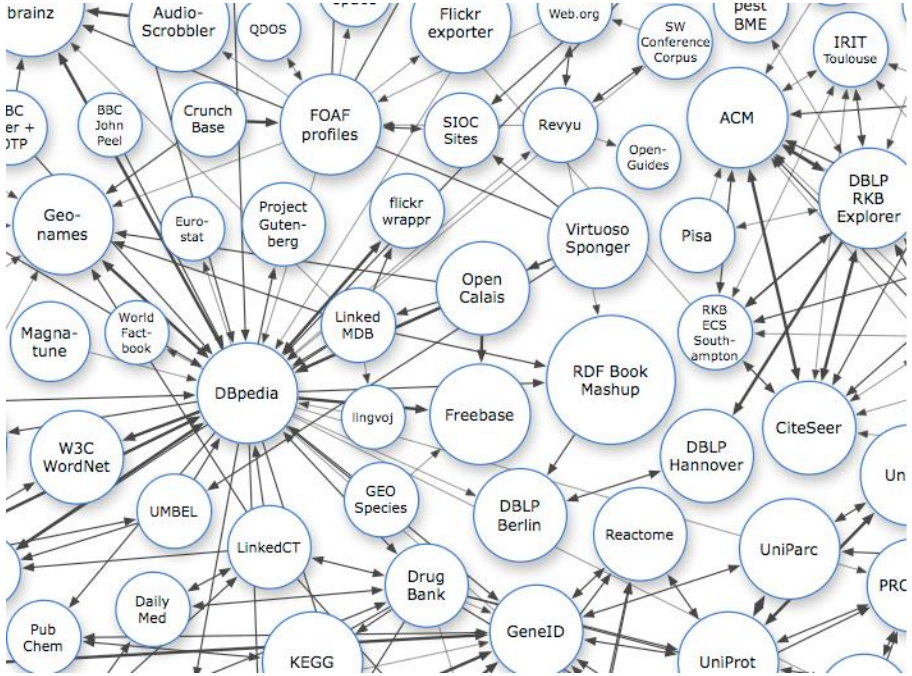
- **What are Examples of KGs?**
  *Freebase, DBPedia, Microsoft Satori, Google Knowledge Graph, Facebook Open Graph, many more*

- **How Large Are KGs?**
  *> 500M entities (topics), 20B facts*

- **How Broad?**
  *Freebase Topics: "American Football" ←→ "Zoos"*

- **What is the Important Characteristic for SLU?**
  - *Web-scale ontology (www.schema.org June 2011)*

➔ ***Massive source of organized NL surface forms***



# Leveraging KGs for Semantic Parsing

## *Procedure*

Unsupervised Data Mining with Knowledge Graphs

- 6 step procedure
- Auto-annotated (unsupervised) data used to train SLU

Style Adaptation

Modeling Relations for Semantic Parsing

# Leveraging KGs for Semantic Parsing

## *Procedure*

Unsupervised Data Mining with Knowledge Graphs

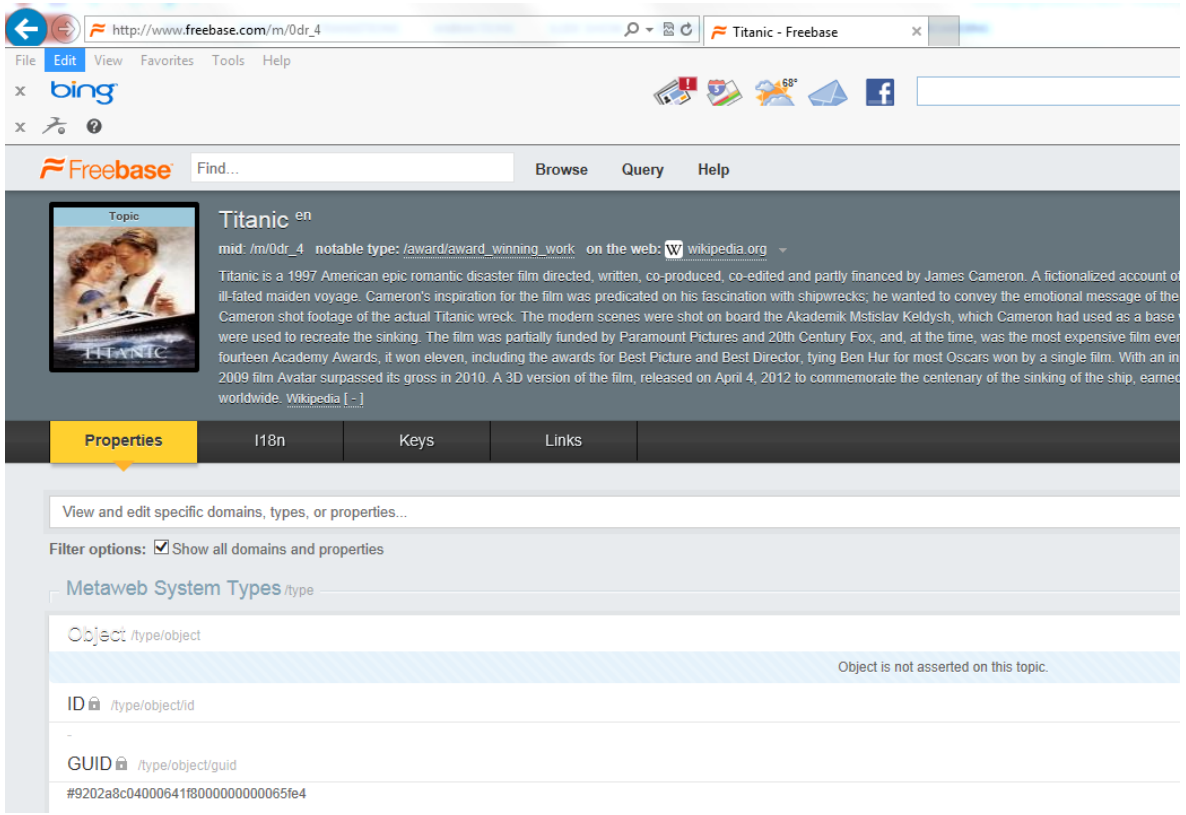
- 6 step procedure
- Auto-annotated (unsupervised) data used to train SLU

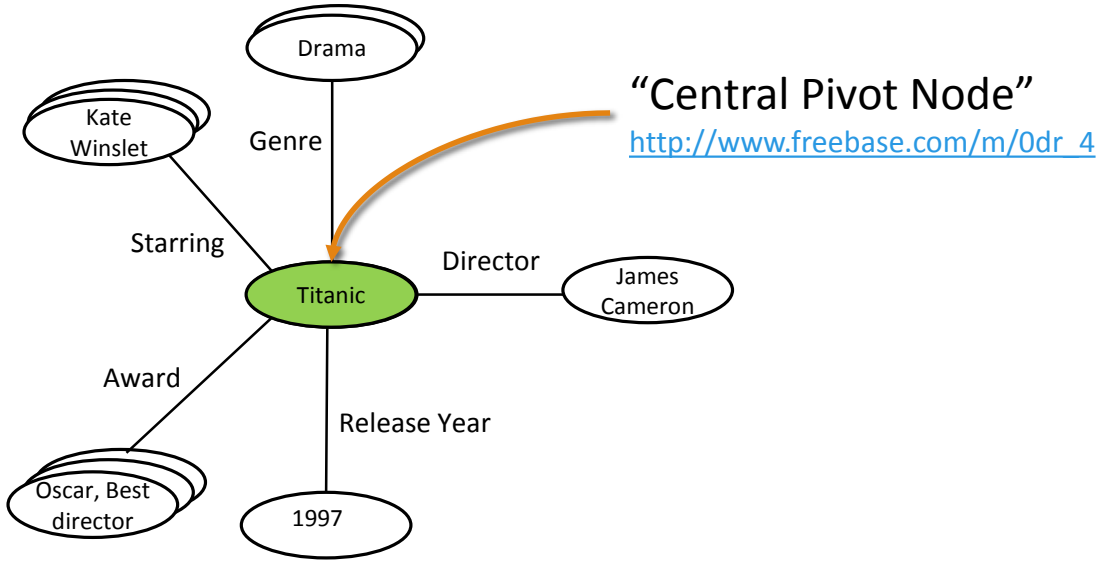
Style Adaptation

Modeling Relations for Semantic Parsing

# Unsupervised Data Mining with KGs

## *Step #1: Select starting node in graph (entity type)*



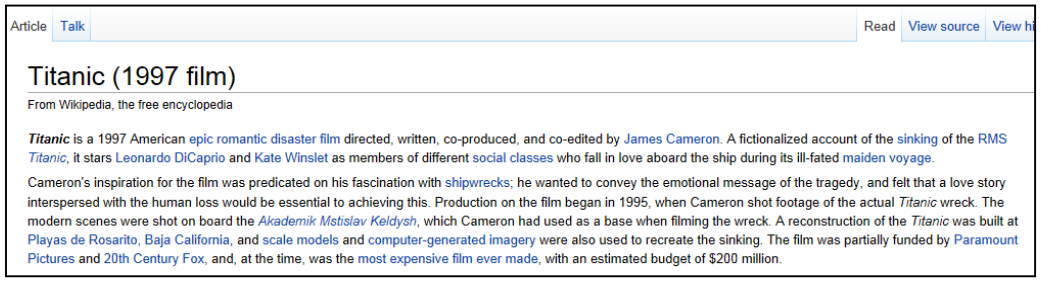



# Unsupervised Data Mining with KGs

## *Step #2: Get Sources of NL Surface Forms*

Freebase links entities to NL Surface forms:

- Wikipedia
- MusicBrainz
- IMDB
- And many more…



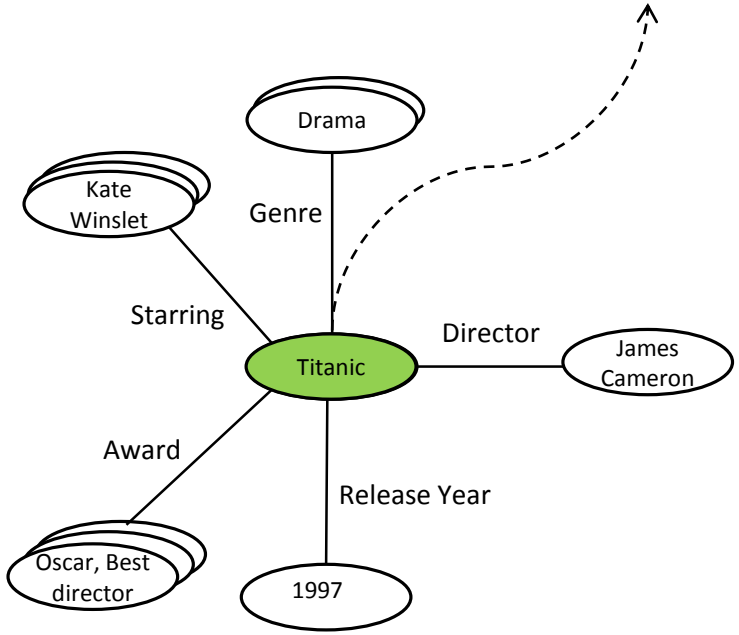
Article Talk Read View source View hi

Titanic (1997 film)

From Wikipedia, the free encyclopedia

***Titanic*** is a 1997 American epic romantic disaster film directed, written, co-produced, and co-edited by James Cameron. A fictionalized account of the sinking of the RMS *Titanic*, it stars Leonardo DiCaprio and Kate Winslet as members of different social classes who fall in love aboard the ship during its ill-fated maiden voyage.

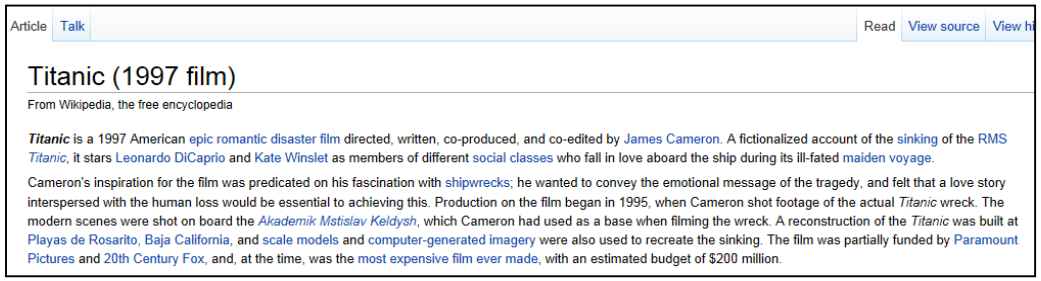
Cameron's inspiration for the film was predicated on his fascination with shipwrecks; he wanted to convey the emotional message of the tragedy, and felt that a love story interspersed with the human loss would be essential to achieving this. Production on the film began in 1995, when Cameron shot footage of the actual *Titanic* wreck. The modern scenes were shot on board the *Akademik Mstislav Keldysh*, which Cameron had used as a base when filming the wreck. A reconstruction of the *Titanic* was built at Playas de Rosarito, Baja California, and scale models and computer-generated imagery were also used to recreate the sinking. The film was partially funded by Paramount Pictures and 20th Century Fox, and, at the time, was the most expensive film ever made, with an estimated budget of $200 million.



# Unsupervised Data Mining with KGs

## *Step #3: Annotate with 1st Order Relations*

| | |
|---|---|
| *Titanic* | B-film_name |
| *stars* | O |
| *Leonardo* | B-film_starring |
| *Dicaprio* | I-film_starring |
| *and* | O |
| *Kate* | B-film_starring |
| *Winslet* | I-film_starring |
| *as* | O |
| … | … |



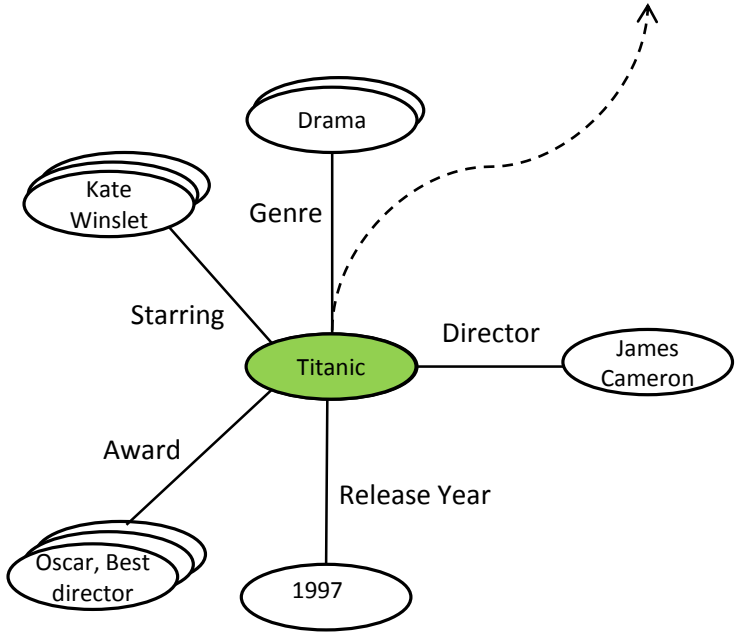
Article Talk Read View source View hi

Titanic (1997 film)

From Wikipedia, the free encyclopedia

***Titanic*** is a 1997 American epic romantic disaster film directed, written, co-produced, and co-edited by James Cameron. A fictionalized account of the sinking of the RMS *Titanic*, it stars Leonardo DiCaprio and Kate Winslet as members of different social classes who fall in love aboard the ship during its ill-fated maiden voyage.

Cameron's inspiration for the film was predicated on his fascination with shipwrecks; he wanted to convey the emotional message of the tragedy, and felt that a love story interspersed with the human loss would be essential to achieving this. Production on the film began in 1995, when Cameron shot footage of the actual *Titanic* wreck. The modern scenes were shot on board the *Akademik Mstislav Keldysh*, which Cameron had used as a base when filming the wreck. A reconstruction of the *Titanic* was built at Playas de Rosarito, Baja California, and scale models and computer-generated imagery were also used to recreate the sinking. The film was partially funded by Paramount Pictures and 20th Century Fox, and, at the time, was the most expensive film ever made, with an estimated budget of $200 million.



# Unsupervised Data Mining with KGs

## *Step #4: Instantiate All Entities of CPN Type*

Article Talk Read Edit source View hi

*Quills*

From Wikipedia, the free encyclopedia

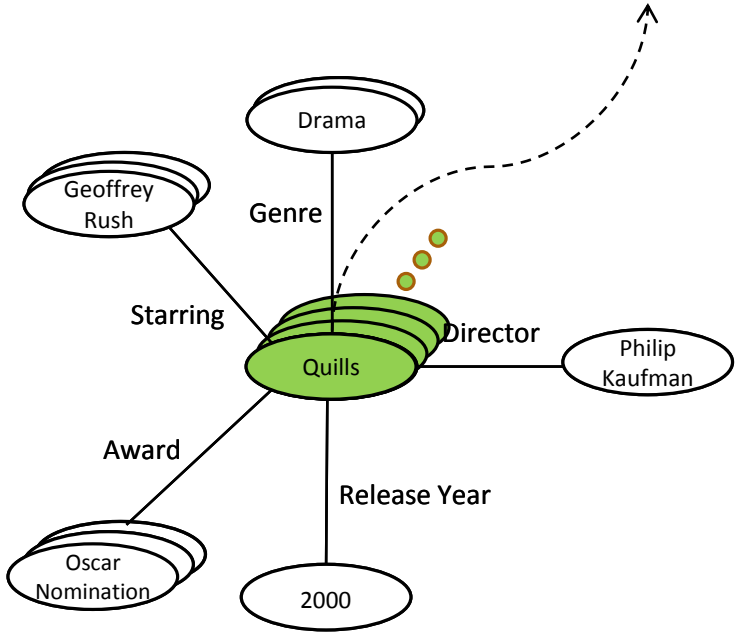
*This article is about the film. For the writing utensils, see Quill. For other uses, see Quill (disambiguation).*

***Quills*** is a 2000 period film directed by Philip Kaufman and adapted from the Obie award-winning play by Doug Wright, who also wrote the original screenplay. Inspired by the life and work of the Marquis de Sade, *Quills* re-imagines the last years of the Marquis' incarceration in the insane asylum at Charenton. It stars Geoffrey Rush as the Marquis de Sade, Joaquin Phoenix as the Abbé du Coulmier, Michael Caine as Dr. Royer-Collard, and Kate Winslet as laundress Madeleine "Maddie" LeClerc.

Well received by critics, *Quills* garnered numerous accolades for Rush, including nominations for an Oscar and a Golden Globe. The film was a modest art house success, averaging $27,709 per screen its debut weekend, and eventually grossing $17,989,277 internationally. Cited by historians as factually inaccurate, *Quills* filmmakers and writers said they were

Explore "depth" of entity-type

→ large entity lists (gazetteers)



# Unsupervised Data Mining with KGs

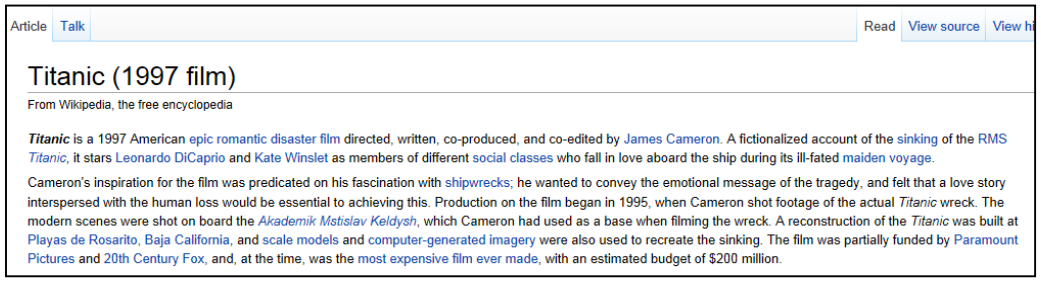
## *Step #5: Get 2nd Order Relations*

Knowledge graph "compositionality"

- Entity-relation templates (grammars) can be composed

| Template | Frequency |
|---|---|
| $ent$ | 44.9% |
| $type \sqcap rel(ent)$ | 12.8% |
| $ent_0 \sqcap rel(ent_1)$ | 7.7% |
| $ent \sqcap type$ | 5.8% |
| $type$ | 5.8% |
| $attr(ent)$ | 3.8% |
| $ent_1 \sqcap rel(ent_0)$ | 3.2% |
| $rel(ent)$ | 1.9% |
| $ent_0 \sqcap rel(ent_1, rel(ent_2))$ | 1.3% |
| $type_1 \sqcap rel(type_0)$ | 1.3% |

Ten most frequently occurring templates among entity-based queries (Pound et al., CIKM'12)



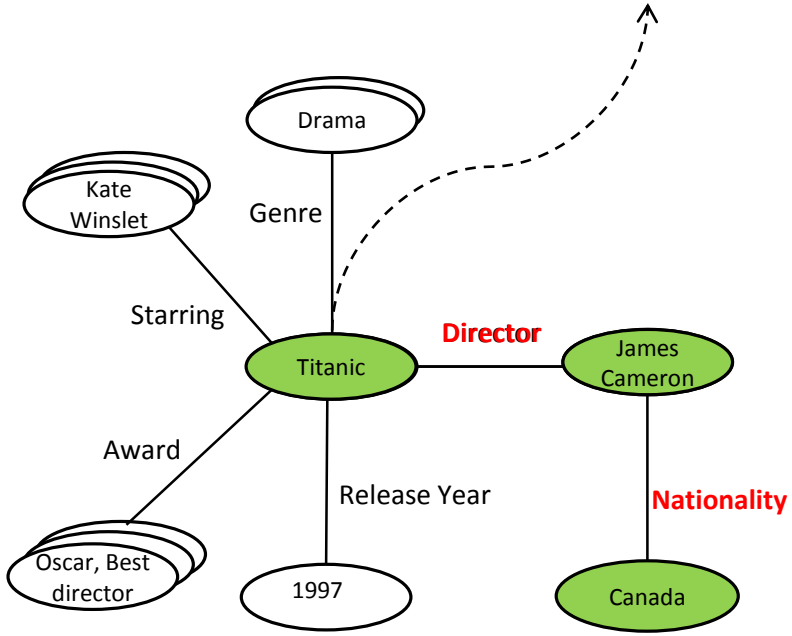
Article Talk Read View source View hi

Titanic (1997 film)

From Wikipedia, the free encyclopedia

***Titanic*** is a 1997 American epic romantic disaster film directed, written, co-produced, and co-edited by James Cameron. A fictionalized account of the sinking of the RMS *Titanic*, it stars Leonardo DiCaprio and Kate Winslet as members of different social classes who fall in love aboard the ship during its ill-fated maiden voyage.

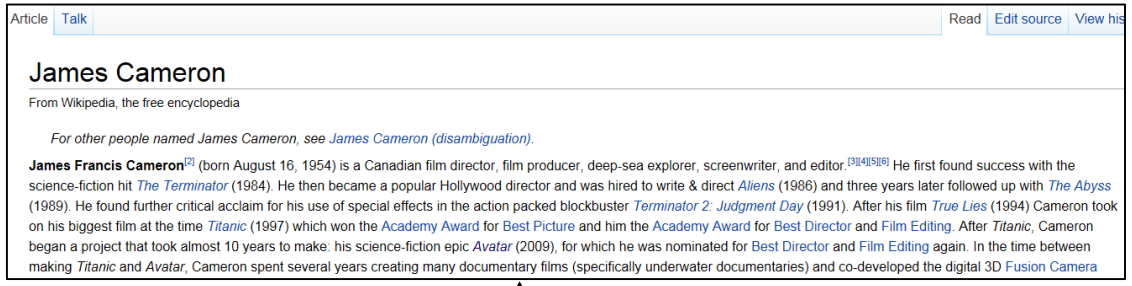
Cameron's inspiration for the film was predicated on his fascination with shipwrecks; he wanted to convey the emotional message of the tragedy, and felt that a love story interspersed with the human loss would be essential to achieving this. Production on the film began in 1995, when Cameron shot footage of the actual *Titanic* wreck. The modern scenes were shot on board the *Akademik Mstislav Keldysh*, which Cameron had used as a base when filming the wreck. A reconstruction of the *Titanic* was built at Playas de Rosarito, Baja California, and scale models and computer-generated imagery were also used to recreate the sinking. The film was partially funded by Paramount Pictures and 20th Century Fox, and, at the time, was the most expensive film ever made, with an estimated budget of $200 million.



# Unsupervised Data Mining with KGs

## *Step #6: Select New CPN and Repeat (Crawl Graph)*

- Select a new central pivot nod
- Repeat steps #1-5
- Crawl the graph until complete



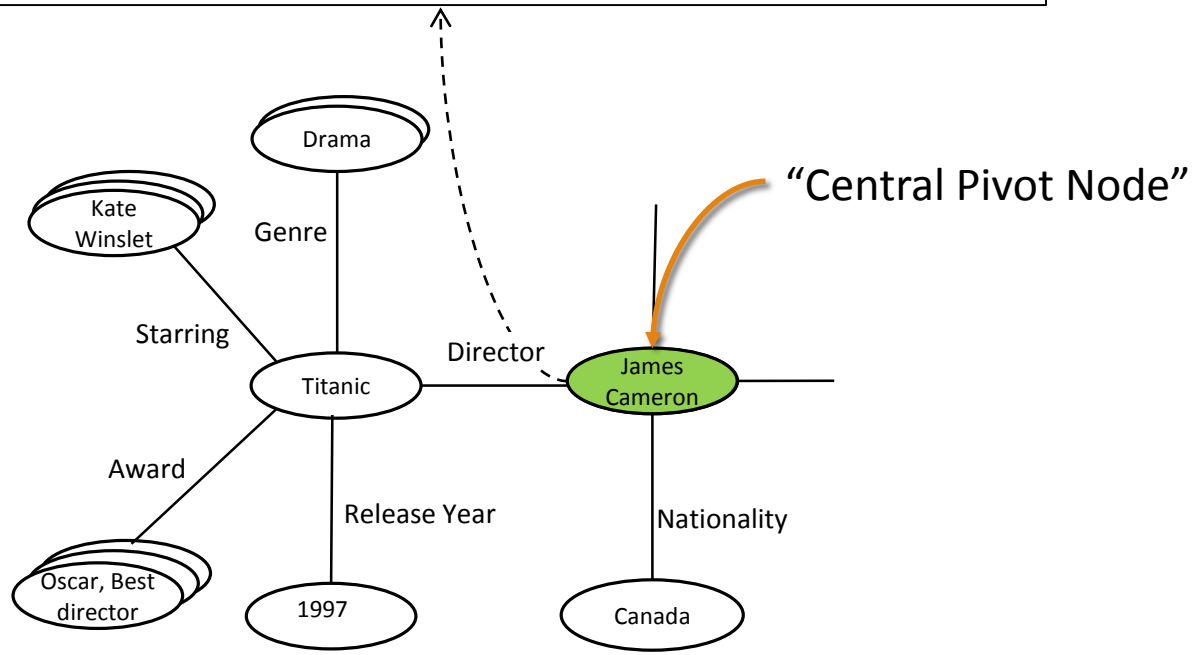
Article Talk Read Edit source View his

James Cameron

From Wikipedia, the free encyclopedia

*For other people named James Cameron, see James Cameron (disambiguation).*

**James Francis Cameron**[2] (born August 16, 1954) is a Canadian film director, film producer, deep-sea explorer, screenwriter, and editor.[3][4][5][6] He first found success with the science-fiction hit *The Terminator* (1984). He then became a popular Hollywood director and was hired to write & direct *Aliens* (1986) and three years later followed up with *The Abyss* (1989). He found further critical acclaim for his use of special effects in the action packed blockbuster *Terminator 2: Judgment Day* (1991). After his film *True Lies* (1994) Cameron took on his biggest film at the time *Titanic* (1997) which won the Academy Award for Best Picture and him the Academy Award for Best Director and Film Editing. After *Titanic*, Cameron began a project that took almost 10 years to make: his science-fiction epic *Avatar* (2009), for which he was nominated for Best Director and Film Editing again. In the time between making *Titanic* and *Avatar*, Cameron spent several years creating many documentary films (specifically underwater documentaries) and co-developed the digital 3D Fusion Camera



# Experimental Setup

Scenario: developer seeks to train a SLU system for a NL movie search application (Netflix)

Training

- Freebase film (movies) domain, 56 relations with linked Wikipedia articles
- Focused on 4 Netflix properties: movies (175K), actors (234K), genres (685), directors (59K)
- 10K NL surface forms = Wikipedia ("Meg Ryan starred with Tom Hanks in …")

Testing

- 2 Conditions
  - Mined Testset
    - Development corpus
    - 1K Wikipedia sentences
    - "Matched" condition
  - Control Testset
    - Target Netflix (true) testset
    - 2K utterances from user data collection

# Results

| | Manual Transcriptions | | | | | ASR Output | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Movie | Actor | Genre | Director | All | Movie | Actor | Genre | Director | All |
| **Supervised** | | | | | | | | | | |
| CRF Lexical + Gazetteers | 51.25% | 86.29% | 93.26% | 64.86% | 66.53% | 45.15% | 82.56% | 88.58% | 58.59% | 60.96% |
| CRF Lexical only | 46.44% | 80.22% | 92.83% | 52.94% | 61.72% | 39.21% | 74.86% | 86.21% | 45.36% | 54.10% |
| **Unsupervised** | | | | | | | | | | |
| Gazetteers only | 69.69% | 50.70% | 15.76% | 2.63% | 51.14% | 59.66% | 47.78% | 11.80% | 2.82% | 43.88% |
| CRF Lexical only | 0.19% | 9.67% | 0.00% | 62.83% | 5.61% | 0.20% | 9.67% | 0.00% | 57.14% | 5.27% |
| + Gazetteers | 1.96% | 72.35% | 4.73% | 79.03% | 31.94% | 1.74% | 69.76% | 3.57% | 75.00% | 30.77% |

Mismatched Style of training (Wikipedia) and testing (Netflix) significantly impacting results

# Leveraging KGs for Semantic Parsing

## *Procedure*

- Unsupervised Data Mining with Knowledge Graphs
  - 6 step procedure
  - Auto-annotated (unsupervised) data used to train SLU
- Style Adaptation
- Modeling Relations for Semantic Parsing

# Adaptation

## *Addressing Mismatch Problem*

Mismatch between training/testing can occur if:

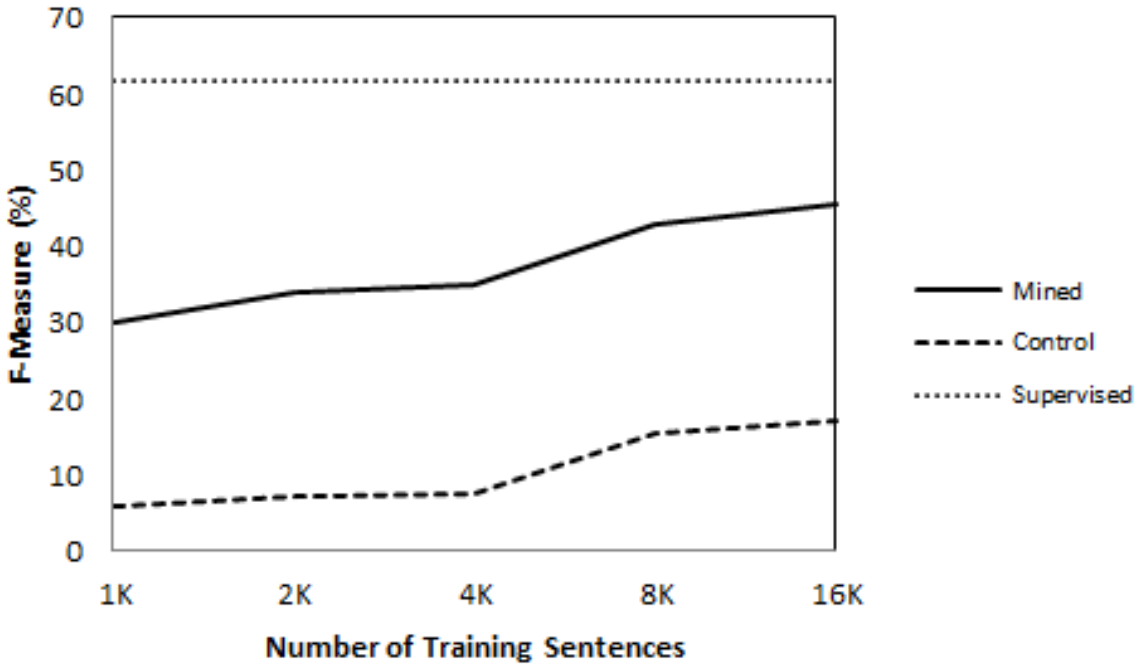
- **Genre Differences**: NL surface forms from knowledge graph sources *mismatch* with the target genre
- **Poor Coverage**: sparse set of surface forms for tail patterns

Results: Freebase + Wikipedia

- **Training:** single source of NL surface forms = Wikipedia ("Meg Ryan starred with Tom Hanks in …")
- **Testing:** Netflix movie search ("show me some funny flicks with Meg Ryan")

Solution

- Rely on relative robustness to mismatch of Gazetteers
- Unsupervised MAP-like bootstrap/retraining adaptation
- Adapt to representative sample of data from target domain




Mined = Wikipedia-Wikipedia
Control = Wikipedia-Netflix

# Results

| | Manual Transcriptions | | | | | ASR Output | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Movie | Actor | Genre | Director | All | Movie | Actor | Genre | Director | All |
| **Supervised** | | | | | | | | | | |
| CRF Lexical + Gazetteers | 51.25% | 86.29% | 93.26% | 64.86% | 66.53% | 45.15% | 82.56% | 88.58% | 58.59% | 60.96% |
| CRF Lexical only | 46.44% | 80.22% | 92.83% | 52.94% | 61.72% | 39.21% | 74.86% | 86.21% | 45.36% | 54.10% |
| **Unsupervised** | | | | | | | | | | |
| Gazetteers only | 69.69% | 50.70% | 15.76% | 2.63% | 51.14% | 59.66% | 47.78% | 11.80% | 2.82% | 43.88% |
| CRF Lexical only | 0.19% | 9.67% | 0.00% | 62.83% | 5.61% | 0.20% | 9.67% | 0.00% | 57.14% | 5.27% |
| + Gazetteers | 1.96% | 72.35% | 4.73% | 79.03% | 31.94% | 1.74% | 69.76% | 3.57% | 75.00% | 30.77% |
| + Adaptation | 71.72% | 58.61% | 29.55% | 77.42% | 60.38% | 55.74% | 62.70% | 30.95% | 73.21% | 54.69% |

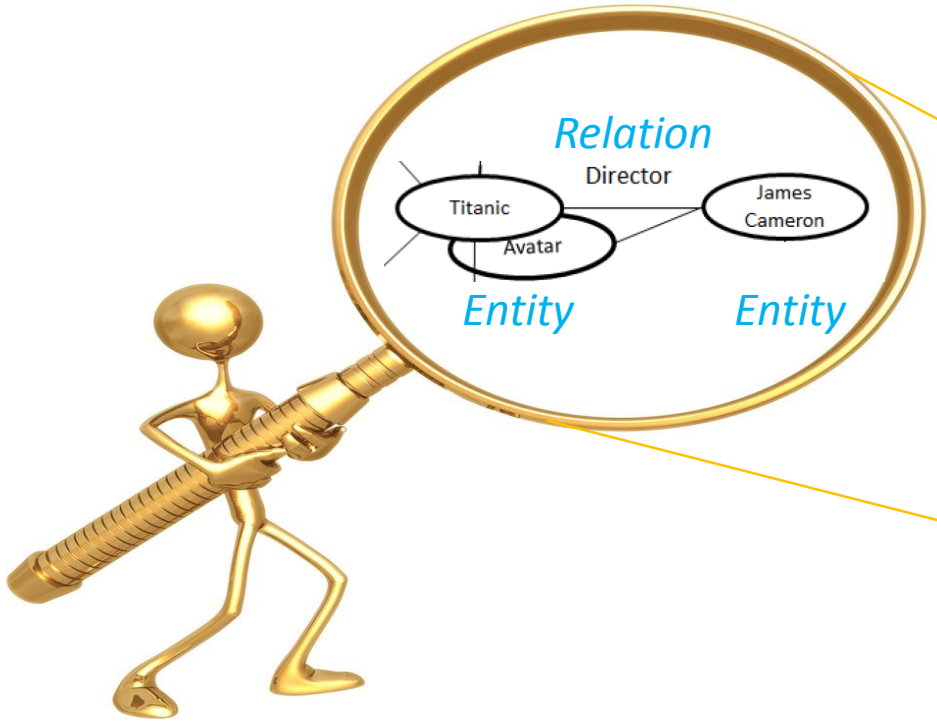
# Leveraging KGs for Semantic Parsing

## *Procedure*

- Unsupervised Data Mining with Knowledge Graphs
  - 6 step procedure
  - Auto-annotated (unsupervised) data used to train SLU
- Style Adaptation
- Modeling Relations for Semantic Parsing

# Modeling *Relations* for Semantic Parsing

## *Semantic Templates*



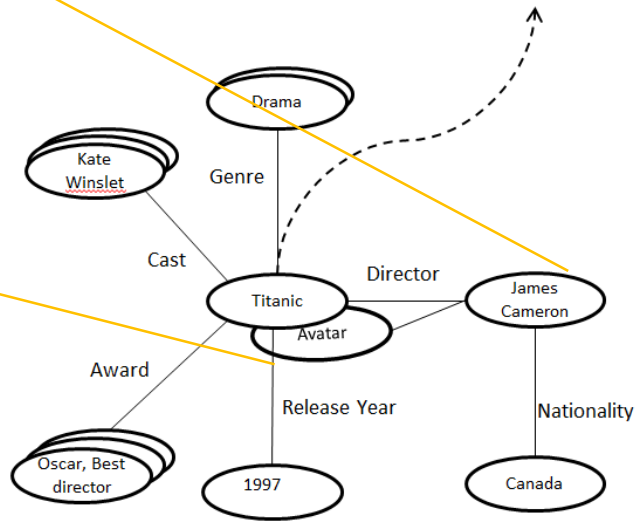

Article Talk Read View source View hi

Titanic (1997 film)

From Wikipedia, the free encyclopedia

***Titanic*** is a 1997 American epic romantic disaster film directed, written, co-produced, and co-edited by James Cameron. A fictionalized account of the sinking of the RMS *Titanic*, it stars Leonardo DiCaprio and Kate Winslet as members of different social classes who fall in love aboard the ship during its ill-fated maiden voyage.

Cameron's inspiration for the film was predicated on his fascination with shipwrecks; he wanted to convey the emotional message of the tragedy, and felt that a love story interspersed with the human loss would be essential to achieving this. Production on the film began in 1995, when Cameron shot footage of the actual *Titanic* wreck. The modern scenes were shot on board the *Akademik Matislav Keldysh*, which Cameron had used as a base when filming the wreck. A reconstruction of the *Titanic* was built at Playas de Rosarito, Baja California, and scale models and computer-generated imagery were also used to recreate the sinking. The film was partially funded by Paramount Pictures and 20th Century Fox, and, at the time, was the most expensive film ever made, with an estimated budget of $200 million.



# Modeling Relations for Semantic Parsing

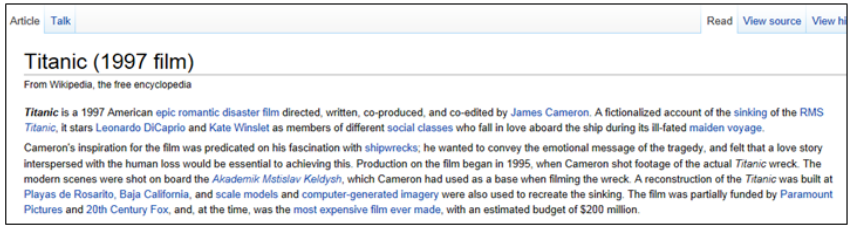
## *Semantic Templates*

Extracted entities provide foundation for higher-level (grammatical) structure

- Leverage our prior work* to identify *entity-relation patterns*
- Induce grammars from templates
- "Repair" missing entities (e.g., "show me *movies with* _____")

| Template | Frequency |
|---|---|
| $ent$ | 44.9% |
| $type \sqcap rel(ent)$ | 12.8% |
| $ent_0 \sqcap rel(ent_1)$ | 7.7% |
| $ent \sqcap type$ | 5.8% |
| $type$ | 5.8% |
| $attr(ent)$ | 3.8% |
| $ent_1 \sqcap rel(ent_0)$ | 3.2% |
| $rel(ent)$ | 1.9% |
| $ent_0 \sqcap rel(ent_1, rel(ent_2))$ | 1.3% |
| $type_1 \sqcap rel(type_0)$ | 1.3% |

Ten most frequently occurring templates among entity-based queries (Pound et al., CIKM'12)

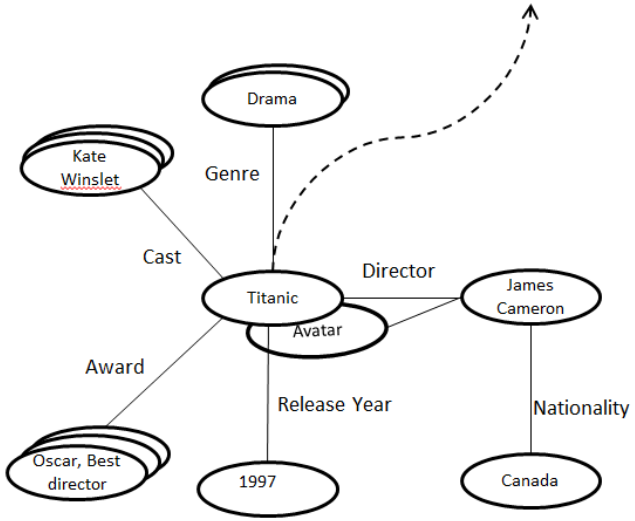

Article Talk Read View source View hi

Titanic (1997 film)

From Wikipedia, the free encyclopedia

***Titanic*** is a 1997 American epic romantic disaster film directed, written, co-produced, and co-edited by James Cameron. A fictionalized account of the sinking of the RMS *Titanic*, it stars Leonardo DiCaprio and Kate Winslet as members of different social classes who fall in love aboard the ship during its ill-fated maiden voyage.

Cameron's inspiration for the film was predicated on his fascination with shipwrecks; he wanted to convey the emotional message of the tragedy, and felt that a love story interspersed with the human loss would be essential to achieving this. Production on the film began in 1995, when Cameron shot footage of the actual *Titanic* wreck. The modern scenes were shot on board the *Akademik Mstislav Keldysh*, which Cameron had used as a base when filming the wreck. A reconstruction of the *Titanic* was built at Playas de Rosarito, Baja California, and scale models and computer-generated imagery were also used to recreate the sinking. The film was partially funded by Paramount Pictures and 20th Century Fox, and, at the time, was the most expensive film ever made, with an estimated budget of $200 million.




* Dilek Hakkani-Tur, Larry Heck, and Gokhan Tur, Using a Knowledge Graph and Query Click Logs for Unsupervised Learning of Relation Detection, ICASSP 2013

# Results

## *Summary*

| | Manual Transcriptions | | | | | ASR Output | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Movie | Actor | Genre | Director | All | Movie | Actor | Genre | Director | All |
| **Supervised** | | | | | | | | | | |
| CRF Lexical + Gazetteers | 51.25% | 86.29% | 93.26% | 64.86% | 66.53% | 45.15% | 82.56% | 88.58% | 58.59% | 60.96% |
| CRF Lexical only | 46.44% | 80.22% | 92.83% | 52.94% | 61.72% | 39.21% | 74.86% | 86.21% | 45.36% | 54.10% |
| **Unsupervised** | | | | | | | | | | |
| Gazetteers only | 69.69% | 50.70% | 15.76% | 2.63% | 51.14% | 59.66% | 47.78% | 11.80% | 2.82% | 43.88% |
| CRF Lexical only | 0.19% | 9.67% | 0.00% | 62.83% | 5.61% | 0.20% | 9.67% | 0.00% | 57.14% | 5.27% |
| + Gazetteers | 1.96% | 72.35% | 4.73% | 79.03% | 31.94% | 1.74% | 69.76% | 3.57% | 75.00% | 30.77% |
| + Adaptation | 71.72% | 58.61% | 29.55% | 77.42% | 60.38% | 55.74% | 62.70% | 30.95% | 73.21% | 54.69% |
| + Relations | | | | **84.62%** | **61.02%** | | | | **80.67%** | **55.40%** |

# Summary

**New approach for unsupervised semantic parsing with knowledge graphs (KGs)**

**Knowledge as Priors:** Leverage large KGs (Freebase) to bootstrap web-scale semantic parsers

- No semantic schema design
- No data collection
- No manual annotations

**Graph Crawling Algorithm for Unsupervised Data Mining**

**Entity and Relation Modeling with Mined Data**

- 61.02% and 55.40% F-measure (Manual/ASR transcriptions)
- Within ***5.5% of supervised training***
- Induced grammars of entity-relation patterns ***increases F-measure by more than 7% absolute*** (Director)