

Natural Language Processing and the Semantic Web

Stuart Robinson Ascander Dost

Powerset, a Microsoft Company

January 28, 2010



Roadmap

Overview

Semantic Search

Conclusion



The Trouble with Triples



- ▶ most of the web is free text, which is difficult to turn into structured data (costly, time-consuming, and error prone)
- ▶ structured data is difficult for non-experts to query and interpret



The Trouble with Triples



- ▶ most of the web is free text, which is difficult to turn into structured data (costly, time-consuming, and error prone)
- ▶ structured data is difficult for non-experts to query and interpret



Trivial Pursuit: Answers For Informational Queries

- ▶ population of sweden
- ▶ how tall is taylor swift?
- ▶ john lennon's children
- ▶ how did tim russert die
- ▶ cast of how the west was won
- ▶ clint eastwood movies



Trivial Pursuit: Answers For Informational Queries

- ▶ population of sweden
- ▶ how tall is taylor swift?
- ▶ john lennon's children
- ▶ how did tim russert die
- ▶ cast of how the west was won
- ▶ clint eastwood movies



Trivial Pursuit: Answers For Informational Queries

- ▶ population of sweden
- ▶ how tall is taylor swift?
- ▶ john lennon's children
- ▶ how did tim russert die
- ▶ cast of how the west was won
- ▶ clint eastwood movies



Trivial Pursuit: Answers For Informational Queries

- ▶ population of sweden
- ▶ how tall is taylor swift?
- ▶ john lennon's children
- ▶ how did tim russert die
- ▶ cast of how the west was won
- ▶ clint eastwood movies



Trivial Pursuit: Answers For Informational Queries

- ▶ population of sweden
- ▶ how tall is taylor swift?
- ▶ john lennon's children
- ▶ how did tim russert die
- ▶ cast of how the west was won
- ▶ clint eastwood movies



Trivial Pursuit: Answers For Informational Queries

- ▶ population of sweden
- ▶ how tall is taylor swift?
- ▶ john lennon's children
- ▶ how did tim russert die
- ▶ cast of how the west was won
- ▶ clint eastwood movies



Freebase

- ▶ User generated structured database of world knowledge
- ▶ Extensive ontology with over 1500 types with 500+ instances
- ▶ Available in multiple formats (including RDF)



Freebase

- ▶ User generated structured database of world knowledge
- ▶ Extensive ontology with over 1500 types with 500+ instances
- ▶ Available in multiple formats (including RDF)

Freebase

- ▶ User generated structured database of world knowledge
- ▶ Extensive ontology with over 1500 types with 500+ instances
- ▶ Available in multiple formats (including RDF)

Querying Freebase

```
clint eastwood movies  =>  [{
    "name" : "clint eastwood",
    "type" : "/film/actor",
    "film" : [{
        "film" : null
    }]
}]
```



Converting User Queries to Instant Answers

- ▶ Extract info from user query to construct structured query (MQL)
 - ▶ name
 - ▶ type
 - ▶ attribute
- ▶ Issue query and obtain structured data (JSON) from source (Freebase)
- ▶ Display structured data in human readable format (rendered HTML)



Converting User Queries to Instant Answers

- ▶ Extract info from user query to construct structured query (MQL)
 - ▶ name
 - ▶ type
 - ▶ attribute
- ▶ Issue query and obtain structured data (JSON) from source (Freebase)
- ▶ Display structured data in human readable format (rendered HTML)



Converting User Queries to Instant Answers

- ▶ Extract info from user query to construct structured query (MQL)
 - ▶ name
 - ▶ type
 - ▶ attribute
- ▶ Issue query and obtain structured data (JSON) from source (Freebase)
- ▶ Display structured data in human readable format (rendered HTML)



Converting User Queries to Instant Answers

- ▶ Extract info from user query to construct structured query (MQL)
 - ▶ name
 - ▶ type
 - ▶ attribute
- ▶ Issue query and obtain structured data (JSON) from source (Freebase)
- ▶ Display structured data in human readable format (rendered HTML)



Converting User Queries to Instant Answers

- ▶ Extract info from user query to construct structured query (MQL)
 - ▶ name
 - ▶ type
 - ▶ attribute
- ▶ Issue query and obtain structured data (JSON) from source (Freebase)
- ▶ Display structured data in human readable format (rendered HTML)



Converting User Queries to Instant Answers

- ▶ Extract info from user query to construct structured query (MQL)
 - ▶ name
 - ▶ type
 - ▶ attribute
- ▶ Issue query and obtain structured data (JSON) from source (Freebase)
- ▶ Display structured data in human readable format (rendered HTML)



Mapping User Queries and Database Queries

clint eastwood movies



<name>clint eastwood</name> <type>/film/actor</type> <attribute>film</attribute>
 <name>clint eastwood</name> <type>/film/director</type> <attribute>film</attribute>

<name>clint eastwood</name> <type>/film/actor</type> <attribute>film</attribute>



clint eastwood movies
 movies starring clint eastwood
 what movies did clint eastwood star in?
 clint eastwood's films

...



Mapping User Queries and Database Queries

clint eastwood movies



<name>clint eastwood</name> <type>/film/actor</type> <attribute>film</attribute>
 <name>clint eastwood</name> <type>/film/director</type> <attribute>film</attribute>

<name>clint eastwood</name> <type>/film/actor</type> <attribute>film</attribute>



clint eastwood movies
 movies starring clint eastwood
 what movies did clint eastwood star in?
 clint eastwood's films

...



Variation in Informational Queries

- ▶ clint eastwood movies
- ▶ films clint eastwood
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Variation in Informational Queries

- ▶ **clint eastwood movies**
- ▶ films clint eastwood
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Variation in Informational Queries

- ▶ **clint eastwood movies**
- ▶ **films clint eastwood**
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Variation in Informational Queries

- ▶ clint eastwood movies
- ▶ films clint eastwood
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Variation in Informational Queries

- ▶ clint eastwood movies
- ▶ films clint eastwood
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Variation in Informational Queries

- ▶ clint eastwood movies
- ▶ films clint eastwood
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Variation in Informational Queries

- ▶ clint eastwood movies
- ▶ films clint eastwood
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Variation in Informational Queries

- ▶ clint eastwood movies
- ▶ films clint eastwood
- ▶ films with clint eastwood
- ▶ movies that clint eastwood acted in
- ▶ films starring clint eastwood
- ▶ which films did clint eastwood star in
- ▶ which films star clint eastwood



Defining Linguistic Templates

- ▶ **motion pictures** starring clint eastwood
- ▶ **movies** starring clint eastwood
- ▶ **films** starring clint eastwood

- ▶ films starring **clint eastwood**
- ▶ films starring **brad pitt**
- ▶ films starring **edward norton**

- ▶ **<FILM>** starring **<ACTOR>**



Defining Linguistic Templates

- ▶ **motion pictures** starring clint eastwood
- ▶ **movies** starring clint eastwood
- ▶ **films** starring clint eastwood

- ▶ films starring **clint eastwood**
- ▶ films starring **brad pitt**
- ▶ films starring **edward norton**

- ▶ <FILM> starring <ACTOR>



Defining Linguistic Templates

- ▶ **motion pictures** starring clint eastwood
- ▶ **movies** starring clint eastwood
- ▶ **films** starring clint eastwood

- ▶ films starring **clint eastwood**
- ▶ films starring **brad pitt**
- ▶ films starring **edward norton**

- ▶ **<FILM>** starring **<ACTOR>**



Defining Linguistic Templates

- ▶ **motion pictures** starring clint eastwood
- ▶ **movies** starring clint eastwood
- ▶ **films** starring clint eastwood

- ▶ films starring **clint eastwood**
- ▶ films starring **brad pitt**
- ▶ films starring **edward norton**

- ▶ <FILM> starring <ACTOR>



Defining Linguistic Templates

- ▶ **motion pictures** starring clint eastwood
- ▶ **movies** starring clint eastwood
- ▶ **films** starring clint eastwood

- ▶ films starring **clint eastwood**
- ▶ films starring **brad pitt**
- ▶ films starring **edward norton**

- ▶ <FILM> starring <ACTOR>



Defining Linguistic Templates

- ▶ **motion pictures** starring clint eastwood
- ▶ **movies** starring clint eastwood
- ▶ **films** starring clint eastwood

- ▶ films starring **clint eastwood**
- ▶ films starring **brad pitt**
- ▶ films starring **edward norton**

- ▶ <FILM> starring <ACTOR>



Defining Linguistic Templates

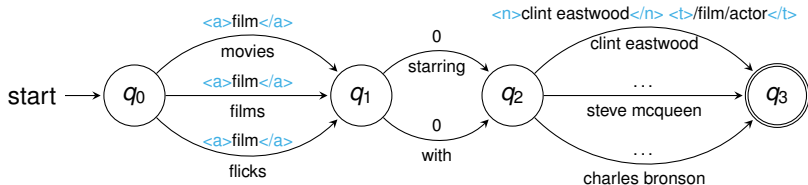
- ▶ **motion pictures** starring clint eastwood
- ▶ **movies** starring clint eastwood
- ▶ **films** starring clint eastwood

- ▶ films starring **clint eastwood**
- ▶ films starring **brad pitt**
- ▶ films starring **edward norton**

- ▶ **<FILM>** starring **<ACTOR>**



An Example Finite-state Transducer



Finite State Transducers (FSTs)

- ▶ finite state transducers are finite state machines with a lower and upper side (two tape)
- ▶ PARC's FST library provides high-level features to handle linguistic phenomena (substitution, insertion, etc.)
- ▶ large scale morphologies available for various languages (Inxight)
- ▶ extensive development and documentation: <http://www.fsmbook.com>



Finite State Transducers (FSTs)

- ▶ finite state transducers are finite state machines with a lower and upper side (two tape)
- ▶ PARC's FST library provides high-level features to handle linguistic phenomena (substitution, insertion, etc.)
- ▶ large scale morphologies available for various languages (Inxight)
- ▶ extensive development and documentation: <http://www.fsmbook.com>



Finite State Transducers (FSTs)

- ▶ finite state transducers are finite state machines with a lower and upper side (two tape)
- ▶ PARC's FST library provides high-level features to handle linguistic phenomena (substitution, insertion, etc.)
- ▶ large scale morphologies available for various languages (Inxight)
- ▶ extensive development and documentation: <http://www.fsmbook.com>



Finite State Transducers (FSTs)

- ▶ finite state transducers are finite state machines with a lower and upper side (two tape)
- ▶ PARC's FST library provides high-level features to handle linguistic phenomena (substitution, insertion, etc.)
- ▶ large scale morphologies available for various languages (Inxight)
- ▶ extensive development and documentation: <http://www.fsmbook.com>



Linguistic Variation in User Queries

- ▶ Morphology: *height* vs. *heights* (singular versus plural)
- ▶ Syntax: *obama height* vs. *height of obama* (keyword versus natural language)
 - ▶ also, variation within natural language
- ▶ Semantics: *film*, *flick*, *celluloid*, *motion picture* (synonymy)
- ▶ Pragmatics: *kennedy's death* (reference)



Linguistic Variation in User Queries

- ▶ Morphology: *height* vs. *heights* (singular versus plural)
- ▶ Syntax: *obama height* vs. *height of obama* (keyword versus natural language)
 - ▶ also, variation within natural language
- ▶ Semantics: *film*, *flick*, *celluloid*, *motion picture* (synonymy)
- ▶ Pragmatics: *kennedy's death* (reference)



Linguistic Variation in User Queries

- ▶ Morphology: *height* vs. *heights* (singular versus plural)
- ▶ Syntax: *obama height* vs. *height of obama* (keyword versus natural language)
 - ▶ also, variation within natural language
- ▶ Semantics: *film*, *flick*, *celluloid*, *motion picture* (synonymy)
- ▶ Pragmatics: *kennedy's death* (reference)



Linguistic Variation in User Queries

- ▶ Morphology: *height* vs. *heights* (singular versus plural)
- ▶ Syntax: *obama height* vs. *height of obama* (keyword versus natural language)
 - ▶ also, variation within natural language
- ▶ Semantics: *film*, *flick*, *celluloid*, *motion picture* (synonymy)
- ▶ Pragmatics: *kennedy's death* (reference)



Linguistic Variation in User Queries

- ▶ Morphology: *height* vs. *heights* (singular versus plural)
- ▶ Syntax: *obama height* vs. *height of obama* (keyword versus natural language)
 - ▶ also, variation within natural language
- ▶ Semantics: *film*, *flick*, *celluloid*, *motion picture* (synonymy)
- ▶ Pragmatics: *kennedy's death* (reference)



Evaluation

- ▶ Apply FSTs to a large query log for evaluation
- ▶ Query log comes from Bing and consists of ≈ 16 M unique queries with frequencies
- ▶ Hand judge each result as good or bad
- ▶ Result: >97% accuracy with 0.06% coverage



Evaluation

- ▶ Apply FSTs to a large query log for evaluation
- ▶ Query log comes from Bing and consists of ≈ 16 M unique queries with frequencies
- ▶ Hand judge each result as good or bad
- ▶ Result: >97% accuracy with 0.06% coverage



Evaluation

- ▶ Apply FSTs to a large query log for evaluation
- ▶ Query log comes from Bing and consists of ≈ 16 M unique queries with frequencies
- ▶ Hand judge each result as good or bad
- ▶ Result: >97% accuracy with 0.06% coverage



Evaluation

- ▶ Apply FSTs to a large query log for evaluation
- ▶ Query log comes from Bing and consists of ≈ 16 M unique queries with frequencies
- ▶ Hand judge each result as good or bad
- ▶ Result: $>97\%$ accuracy with 0.06% coverage



Accuracy and Coverage Figures

| Type | Count | Accuracy | Coverage |
|---------------------------------|-------|----------|----------|
| /film/actor | 3286 | 99.0261 | 0.0205 |
| /people/person | 1872 | 99.1987 | 0.0117 |
| /people/deceased_person | 974 | 99.8973 | 0.0061 |
| /visual_art/art_subject | 886 | 92.8893 | 0.0055 |
| /film/film | 683 | 95.7540 | 0.0042 |
| /film/director | 621 | 97.4235 | 0.0038 |
| /location/statistical_region | 331 | 100.0000 | 0.0020 |
| /business/company | 267 | 71.9101 | 0.0016 |
| /visual_art/visual_artist | 223 | 96.4125 | 0.0013 |
| /visual_art/art_period_movement | 74 | 94.5945 | 0.0004 |
| /business/company_founder | 13 | 76.9230 | 0.0000 |
| /visual_art/artwork | 13 | 46.1538 | 0.0000 |



Accuracy and Coverage Figures

| Type | Count | Accuracy | Coverage |
|---------------------------------|-------|----------|----------|
| /film/actor | 3286 | 99.0261 | 0.0205 |
| /people/person | 1872 | 99.1987 | 0.0117 |
| /people/deceased_person | 974 | 99.8973 | 0.0061 |
| /visual_art/art_subject | 886 | 92.8893 | 0.0055 |
| /film/film | 683 | 95.7540 | 0.0042 |
| /film/director | 621 | 97.4235 | 0.0038 |
| /location/statistical_region | 331 | 100.0000 | 0.0020 |
| /business/company | 267 | 71.9101 | 0.0016 |
| /visual_art/visual_artist | 223 | 96.4125 | 0.0013 |
| /visual_art/art_period_movement | 74 | 94.5945 | 0.0004 |
| /business/company_founder | 13 | 76.9230 | 0.0000 |
| /visual_art/artwork | 13 | 46.1538 | 0.0000 |

Accuracy and Coverage Figures

| Type | Count | Accuracy | Coverage |
|---------------------------------|-------|----------|----------|
| /film/actor | 3286 | 99.0261 | 0.0205 |
| /people/person | 1872 | 99.1987 | 0.0117 |
| /people/deceased_person | 974 | 99.8973 | 0.0061 |
| /visual_art/art_subject | 886 | 92.8893 | 0.0055 |
| /film/film | 683 | 95.7540 | 0.0042 |
| /film/director | 621 | 97.4235 | 0.0038 |
| /location/statistical_region | 331 | 100.0000 | 0.0020 |
| /business/company | 267 | 71.9101 | 0.0016 |
| /visual_art/visual_artist | 223 | 96.4125 | 0.0013 |
| /visual_art/art_period_movement | 74 | 94.5945 | 0.0004 |
| /business/company_founder | 13 | 76.9230 | 0.0000 |
| /visual_art/artwork | 13 | 46.1538 | 0.0000 |



Accuracy and Coverage Figures

| Type | Count | Accuracy | Coverage |
|---------------------------------|-------|----------|----------|
| /film/actor | 3286 | 99.0261 | 0.0205 |
| /people/person | 1872 | 99.1987 | 0.0117 |
| /people/deceased_person | 974 | 99.8973 | 0.0061 |
| /visual_art/art_subject | 886 | 92.8893 | 0.0055 |
| /film/film | 683 | 95.7540 | 0.0042 |
| /film/director | 621 | 97.4235 | 0.0038 |
| /location/statistical_region | 331 | 100.0000 | 0.0020 |
| /business/company | 267 | 71.9101 | 0.0016 |
| /visual_art/visual_artist | 223 | 96.4125 | 0.0013 |
| /visual_art/art_period_movement | 74 | 94.5945 | 0.0004 |
| /business/company_founder | 13 | 76.9230 | 0.0000 |
| /visual_art/artwork | 13 | 46.1538 | 0.0000 |



Bad Mappings

| Query | Name | Type | Attribute |
|--------------------------|--------------|---------------------------|-------------------|
| berkeley heights | berkeley | /architecture/structure | height_meters |
| cost of fireworks | fireworks | /film/film | estimated_budget |
| music for new year | new year | /film/film | music |
| pictures of adolf hitler | adolf hitler | /visual_art/visual_artist | artworks |
| clothing for women | women | /film/film | costume_design_by |
| hero creator | hero | /business/company | founders |



Acknowledgements

- ▶ Instant Answers Team: Chris Jewell, Max Lansing, Andrei Makhanov, Franco Salvetti
- ▶ Text Processing for Semantic Applications, Natural Language Engineering: Richard Crouch, Tracy King, Martin Forst, Olya Gurevich, Martin van den Berg, Scott Waterman
- ▶ FST Experts: Ronald Kaplan, Lauri Karttunen (PARC)



Acknowledgements

- ▶ Instant Answers Team: Chris Jewell, Max Lansing, Andrei Makhanov, Franco Salvetti
- ▶ Text Processing for Semantic Applications, Natural Language Engineering: Richard Crouch, Tracy King, Martin Forst, Olya Gurevich, Martin van den Berg, Scott Waterman
- ▶ FST Experts: Ronald Kaplan, Lauri Karttunen (PARC)



Acknowledgements

- ▶ Instant Answers Team: Chris Jewell, Max Lansing, Andrei Makhanov, Franco Salvetti
- ▶ Text Processing for Semantic Applications, Natural Language Engineering: Richard Crouch, Tracy King, Martin Forst, Olya Gurevich, Martin van den Berg, Scott Waterman
- ▶ FST Experts: Ronald Kaplan, Lauri Karttunen (PARC)



The End

Thanks!