Mayank Kejriwal
University of Southern California

Open Information Extraction

# Application: Information Fusion

- What kills bacteria?
- What west coast, nano-technology companies are hiring?
- Compare Obama's "buzz" versus Hillary's?
- What is a quiet, inexpensive, 4-star hotel in Vancouver?



slide by Oren Etzioni

# Opinion Mining

- Opine **(Popescu & Etzioni, EMNLP '05)**
- IE(product reviews)
  - Informative
  - Abundant, but varied
  - Textual
- Summarize reviews without **any** prior knowledge of product category



slide by Oren Etzioni

# OPINE

Ana-Maria Popescu, Bao Nguyen, Oren Etzioni

Home | Language: English

New York City hotels > Renaissance New York Hotel Times Square

**Review Summary**

**Staff**: excellent (7), great (3), very helpful (2), poor, fantastic, helpful, love, good, *view all* (17)

**Location**: great (4), best (3), good (2), fabulous, fantastic, ideal, superb, not great, love, *view all* (15)

**Room**: nice (5), great (2), not great (2), good (2), very nice (2), excellent, superb, lovely, average, *view all* (17)

**Quality**: best, fantastic, lovely, recommend, love, nice, fine, *view all* (7)

**Food**: very good (2), fantastic, lovely, not great, great, *view all* (6)

**Bathroom beauty**: beautiful

**Bar**: fabulous, great, *view all* (2)

**Staff friendliness**: friendly (4), very friendly (2), incredibly friendly, unfriendly, *view all* (8)

**Room bed comfort**: comfy (2), comfortable (2), extremely comfortable, *view all* (5)

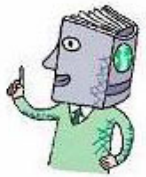
**Bathroom**: great (2), elegant, very nice, nice, *view all* (5)

**Room cleanness**: **clean (2)**

**User comments:**

the rooms were clean and smelled great . Read more

The rooms were clean, spacious, soundproof and well-appointed . Read more



# OPINE

Ana-Maria Popescu, Bao Nguyen, Oren Etzioni

Home | Language: English

New York City hotels > A Greenwich Village Habitue

**Review Summary**

**Canal house beauty**: beautiful

**Location**: perfect

**Room**: gorgeous

**Room cleanness**: spotlessly clean

**City cleanness**: clean

**City comfort**: comfortable

**Room bed comfort**: comfortable

**Bar distance**: close

**Add new opinion**:

Feature: staff Opinion: beautiful Opinion text (optional)

When compared to Renaissance New York Hotel Times Square, **Room cleanness** is

- better at A Greenwich Village Habitue (49 others)
- worse at Morningside Inn (34 others)
- similar at Chelsea Inn - 17th Street (86 others)

**Better hotels:**

New York City hotels > A Greenwich Village Habitue

New York City hotels > Union Square Inn

New York City hotels > Sofitel New York

New York City hotels > Second Home on Second Avenue

New York City hotels > The Muse

New York City hotels > Belleclaire Hotel

New York City hotels > The St. Regis

New York City hotels > Kitano New York

New York City hotels > Milburn Hotel

New York City hotels > Hotel 41 At Times



slide by Oren Etzioni

# Open IE = Self-supervised IE

(Banko, Cafarella, Soderland, et. al, IJCAI '07)

| | Traditional IE | Open IE |
|---|---|---|
| **Input:** | Corpus + Hand-labeled Data | Corpus |
| **Relations:** | Specified in Advance | Discovered Automatically |
| **Complexity:** | $O(D * R)$<br>$R$ relations | $O(D)$<br>$D$ documents |
| **Text analysis:** | Parser + Named-entity tagger | NP Chunker |



slide by Oren Etzioni

# Extractor Overview (Banko & Etzioni, ’08)

1. Use a simple model of relationships in English to label extractions
2. Bootstrap a **general** model of relationships in English sentences, encoded as a CRF
3. Decompose each sentence into one or more (NP1, VP, NP2) “chunks”
4. Use CRF model to retain relevant parts of each NP and VP.

**The extractor is relation-independent!**



slide by Oren Etzioni

# TextRunner Extraction

- Extract Triple representing binary relation **(Arg1, Relation, Arg2)** from sentence.

Internet powerhouse, EBay, was originally founded by Pierre Omidyar**.**

*Internet powerhouse,* **EBay,** *was originally* **founded by Pierre Omidyar.**

**(Ebay, Founded by, Pierre Omidyar)**



slide by Oren Etzioni

# Numerous Extraction Challenges

- **Drop non-essential info:**

"was originally founded by" → **founded by**

- **Retain key distinctions**

**Ebay** founded **by** **Pierre** ≠ **Ebay** founded **Pierre**

- **Non-verb relationships**

"George Bush, president of the U.S…"

- **Synonymy & aliasing**

Albert Einstein = Einstein ≠ Einstein Bros.



slide by Oren Etzioni

# TextRunner (Web's 1st Open IE system)

1. **Self-Supervised Learner**: automatically labels example extractions & learns an extractor

2. **Single-Pass Extractor:** single pass over corpus, identifying extractions in each sentence

3. **Query Processor:** indexes extractions→ enables queries at interactive speeds



slide by Oren Etzioni

# 4. Conclusions

Imagine search systems that operate over a (more) semantic space

- Key words, documents ➔ **extractions**
- TF-IDF, pagerank ➔ **relational models**
- Web pages, hyper links ➔ **entities, relns**

***Reading the Web ➔ new Search Paradigm***



slide by Oren Etzioni

# Open IE on Web Text

| **Advantages** | **Challenges** |
| --- | --- |
| **"Semantically tractable" sentences** | **Difficult, ungrammatical sentences** |
| **Redundancy** | **Unreliable information** |
| **Search engines** | **Heterogeneous corpus** |

slide by Oren Etzioni