

Tackling the Challenges of Big Data Big Data Analytics

Regina Barzilay

Professor

Massachusetts Institute of Technology



© 2014 Massachusetts Institute of Technology

Tackling the Challenges of Big Data Big Data Analytics

Information Summarization

Information Extraction from Social Media

Regina Barzilay

Professor

Massachusetts Institute of Technology



© 2014 Massachusetts Institute of Technology

My Research

- Focus: Natural Language Processing
- Problems:
 - Syntactic analysis (Parsing, Tagging)
 - Multilingual learning
 - Summarization
 - Information Extraction
 - Lost language decipherment



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology

Need in Information Extraction

- Lots of useful data in natural language
 - 4 of 4 people found the following review helpful
 Fits perfect!
By A. Griff on February 2, 2012
Amazon Verified Purchase
I was worried that these one-size-fits-all gloves would be too big, but they fit perfect. I have medium size hands and they are perfect. If you have small to large hands, they should also work great. If you have extra large hands, probably they wont fit.

The material is nice and the stitching seems good as well. They probably wont last forever, but 3 pairs for this price it's a complete deal. I say BUY!

Comment Was this review helpful to you? Yes No
- Commonly used in decision making
- Severely under-utilized

Information Extraction

- **Goal:** Transfer information in natural language into a structured form

```
graph LR; A[Texts] --> B[IE System]; B --> C[Database]
```

The diagram illustrates the flow of information extraction. It consists of three rounded rectangular boxes arranged horizontally. The first box on the left is labeled "Texts". An orange arrow points from "Texts" to the second box, which is labeled "IE System". Another orange arrow points from "IE System" to the third box, which is labeled "Database". All three boxes have a thin red border.

- **Benefits:** Enables to effectively analyze, aggregate and compare the information

The Power of Word Counts

- Simple statistical models are effective for many Information Extraction tasks



- Example:** Named Entity Disambiguation
 - Labels: Person, Organization, Location, Non-entity
 - Features: Capitalization, Belongs to a list of names, Previous word, etc.
 - Classifier learns the mapping from features to labels

Sequence Labeling for Information Extraction

- Assign role labels for words in a sentence

@YonderMountain rocking Mercury Lounge

- Make mutually consistent decisions on a sequence – label assignments are not independent
- Models:** Hidden Markov Models, Conditional Random Fields

Information Extraction for Big Data

- **Challenge:** Supervised setup requires large amounts of training data
 - Typically requires manual annotations
 - Expensive and hard to collect
 - Genre and topic specific
- ➔ Supervised data is small
- **Opportunity:** Use Big Data to augment supervised learning
 - Uncover structured regularities in Big Data, which are relevant for the IE task

Multi-Aspect Summarization

Crags is a great place for the food I much like. The food was great and inexpensive. Pros:
Pros:
I had never been there before. Crags is such a nice place.
I would have never thought of that somehow.
I was very satisfied with the food. It was delicious and fresh.
and then change we eat at the pre-Party.
The Spas. The decor was excellent and
and the food was great.
Cons:
I found it was interesting. Yes, same as the pros.

 Crags has my husband like their leading menu, and we were not disappointed. It was
such a wonderful meal last year and I'm looking down the class until I go back.
The service was polite and knowledgeable, the atmosphere was elegant and energetic,
and the food was wonderfully creative and delicious.

 I like some of the chef's notes which consisted of five meals along the same.
I like the way they are presented, unique and original, making appetizers. Such a delicious job. I don't
think I could ever eat in a sit-down.

 Although the leading menu was widespread (and delicious), we decided to create a little
bit more of our own. We can't wait to eat again, and the staff is great.



Aspect	Snippets
atmosphere	"elegant and energetic" "awesome art"
food	"loved it!" "tasty calzones!"
service	"fast and friendly" "impatient waiters"

Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

The diagram illustrates the process of extracting information from two tweets. On the left, two blue Twitter bird icons are positioned above two rounded rectangular boxes containing text. The top box contains the text: "Seated at @carnegiehall waiting for @CraigyFerg's show". The bottom box contains the text: "RT @leerader : getting REALLY stoked for #CraigyAtCarnegie sat night." To the right of these boxes is a large curly brace spanning both boxes, indicating they belong to the same category. To the right of the brace is a vertical stack of three rectangular boxes. The top box is labeled "Artist:" and contains the name "Craig Ferguson". The middle box is labeled "Venue:" and contains the location "Carnegie Hall".

Tackling the Challenges of Big Data

Big Data Analytics

Information Summarization

Information Extraction from Social Media

Tackling the Challenges of Big Data

Big Data Analytics

Regina Barzilay

Professor

Massachusetts Institute of Technology

Tackling the Challenges of Big Data

Big Data Analytics

Information Summarization

Multi-aspect Summarization

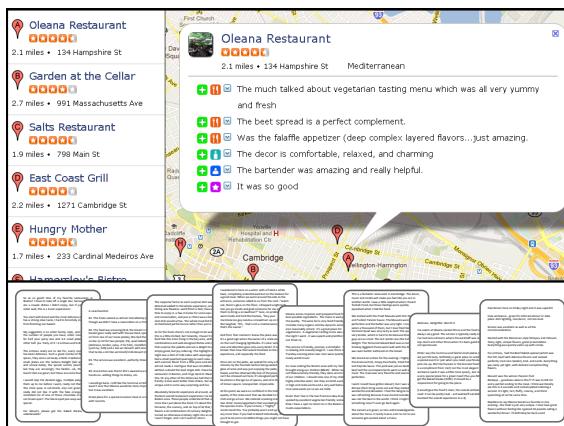
Regina Barzilay

Professor

Massachusetts Institute of Technology



© 2014 Massachusetts Institute of Technology



Multi-Aspect Summarization

Aspect	Snippets
atmosphere	"stylish decor" "awesome art"
food	"loved it!" "tasty calzones!"
service	"fast and friendly" "impatient waiters"

Importance of Context:

... by local artists.
Ordered chicken
parm and loved it!
Friend had the veal.
The service was ...

food {

Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

DATA SCIENCE

Traditional Approach

Task Labels: Observed

I had the shrimp salad and was [FOOD pleasantly surprised]. The [ATMOSPHERE decor was tasteful] and staff was [SERVICE extremely professional and efficient].

MIT PROFESSIONAL EDUCATION Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

Adding Context

Sequence Labeling Task

I ordered lunch from them the other day and I was [FOOD pleasantly surprised]. Our waiter dazzled me with his blue eyes and genuine smile, and all the waiters were [SERVICE extremely professional and efficient].

Content Topic Model

I ordered lunch from them the other day and I was pleasantly surprised. Our waiter dazzled me with his blue eyes and genuine smile, and all the waiters were extremely professional and efficient.

MIT PROFESSIONAL EDUCATION Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

Our Approach

Content Labels: Latent

Task Labels: Observed

Goal: Use global content to assist with local decisions

MIT PROFESSIONAL EDUCATION Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

Multi-Aspect Summarization

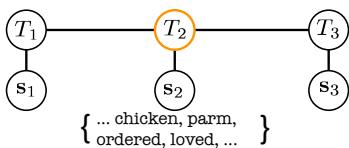
Content Model: Sentence-Level HMM

```

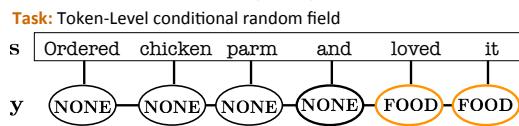
graph LR
    T1((T1)) --- S1((S1))
    T2((T2)) --- S2((S2))
    T3((T3)) --- S3((S3))
    bracket["{ ... chicken, parm, \nordered, loved, ... }"] --- S1
    bracket --- S2
    bracket --- S3
  
```

Task: Token-Level conditional random field

s	Ordered	chicken	parm	and	loved	it
y	(NONE)	(NONE)	(NONE)	(NONE)	FOOD	FOOD



{ ... chicken, parm,
ordered, loved, ... }



PROGRESSIVE
POLITICAL
PARTY

Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology

Factorization

$$P(\mathbf{T}, \mathbf{s}, \mathbf{y}) = P_{\theta}(\mathbf{T}, \mathbf{s}) P_{\phi}(\mathbf{y} | \mathbf{T}, \mathbf{s})$$

$$= \prod_{i=1}^n P_{\theta}(T_{i+1} | T_i) \underbrace{(P_{\theta}(s_i | T_i) P_{\phi}(y_i | s_i, T_i))}_{\text{Topic Trans.}} \underbrace{\quad}_{\text{Bag-of-words}} \underbrace{\quad}_{\text{CRF}}$$

Product over sentences

Topic Trans.

Bag-of-words

CRF

θ Topic Params
 ϕ Task Params

T Sent. Topics
y Task Labels

s Sentences

11.11 | BUSINESS ANALYTICS

MIP
PROFESSIONAL
EDUCATION

Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



- Train: 35 reviews
 - Test: 24 reviews
 - Unlabeled: 12,600 reviews
 - **Yelp restaurant reviews**
 - Train: 48 reviews
 - Test: 48 reviews
 - Unlabeled: 33,000 reviews



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Information Extraction

Goal: Extract phrases from review text in pre-specified categories

Input: User-generated review text, labeled training data

Output: Labeled phrases in each category

Category	Review Text	Extracted Phrases
FOOD	Crusoe is a great place for the barbecue. The ribs were delicious.	the barbecue, ribs
FOOD	Food was Interesting. Crispy prawns itself would have never thought of that somehow.	interesting, crispy prawns
SERVICE	The service was polite and energetic; and their forms change we put in through; the food was wonderful creative and delicious.	polite, energetic, forms change, food, delicious
ATMOSPHERE	-Food was Interesting. Yes, same as the atmosphere.	interesting, atmosphere
PRICE	Although the price was a little bit expensive, we didn't mind it.	price, expensive
OVERALL	I came here with my husband for the tasting menu, and we were not disappointed. We got to sit at the chef's table, which overlooked the kitchen. The service was polite and knowledgeable, the atmosphere was elegant and energetic and the food was wonderfully creative and delicious.	tasting menu, chef's table, service, atmosphere, food

- **NoCM:** Just the CRF, no content model
- **IndepCM:** Estimate content model parameters first, then use them in the CRF
- **JointCM:** Estimate content and CRF parameters jointly using EM

Results

Token F-measure Evaluation

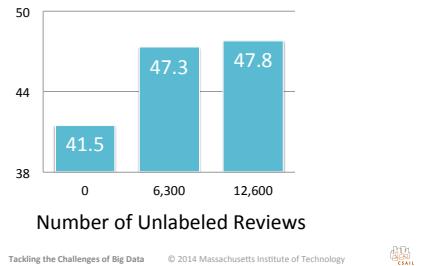
The chart displays the Token F-measure for three models (NoCM, IndepCM, and JointCM) across two datasets (Amazon and Yelp). The Y-axis represents the Token F-measure, ranging from 25 to 50.

Dataset	NoCM	IndepCM	JointCM
Amazon	35	43	47.8
Yelp	28.8	37.9	39.2

Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

Impact of Unlabeled Data

Setup: Using the Amazon corpus, fix the amount of labeled data, vary the amount of unlabeled data



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Tackling the Challenges of Big Data

Big Data Analytics

Information Summarization

Multi-Aspect Summarization

THANK YOU



© 2014 Massachusetts Institute of Technology

Tackling the Challenges of Big Data

Big Data Analytics

Regina Barzilay

Professor

Massachusetts Institute of Technology



© 2014 Massachusetts Institute of Technology

Tackling the Challenges of Big Data

Big Data Analytics

Information Summarization

Event Discovery in Social Media Feeds

Regina Barzilay

Professor

Massachusetts Institute of Technology



© 2014 Massachusetts Institute of Technology

The Task

- Goal: Automatic construction of event records from Twitter
- Input: Stream of Twitter messages
 - Seated at @carnegiehall waiting for @CraigyFerg's show
 - @DJPaulyD absolutely killed it at Terminal 5 last night.
 - Craig, nice seeing you #noelnight this weekend @becksdavis!
- Output: Table of event records

Artist	Venue
Craig Ferguson	Carnegie Hall
DJ Pauly D	Terminal 5



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



IE for Social Media: Challenges

- Messages are short
 - ⇒ Individual message may not contain all event fields.
- Message are expressed in colloquial language
 - ⇒ Mapping between messages and event record is not obvious

Seated at @carnegiehall waiting
for @CraigyFerg's show

RT @leerader : getting REALLY
stoked for #CraigyAtCarnegie
sat night.

Artist: Craig Ferguson
Venue: Carnegie Hall



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



IE for Social Media: Opportunity

Significant redundancy in Twitter stream:

```
Seated at @carnegiehall waiting for @CraigyFerg's show
@DJPaulyD absolutely killed it at Terminal 5 last night.
Craig, nice seeing you #noelnight this weekend @becksdavis!
```

Approach: Drive event extraction by modeling agreement in message stream.



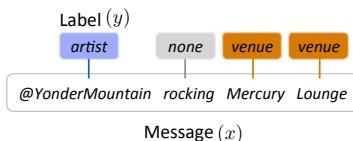
Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Model Functionality

- Message level analysis: Tag words in message with event-field labels.



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



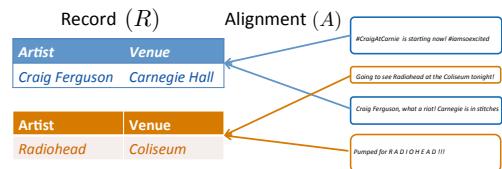
Model Functionality

- Message level analysis: Tag words in message with event-field labels.
- Message clustering: Group messages based on events.
- Event records: Induce canonical value for each field.



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Model Overview

Source of supervision: Example event records

- Alignment between records and messages not observed.
- Message level field annotations not observed.



Tackling the Challenges of Big Data

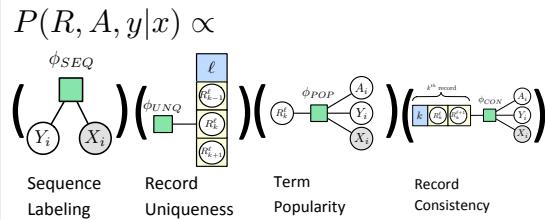
© 2014 Massachusetts Institute of Technology



Model Overview

- (y) Message level analysis
- (A) Message clustering
- (R) Event records

Learn jointly in factor graph model

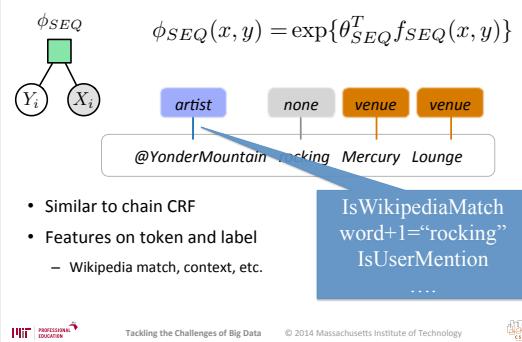


Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Sequence Labeling Factor

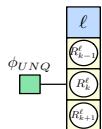


Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Record Uniqueness Factor



$$\phi_{UNQ}(\mathbf{R}^\ell) = \prod_{k \neq k'} \phi_{UNQ}(R_k^\ell, R_{k'}^\ell)$$

$$\phi_{UNQ}(R_k^\ell, R_{k'}^\ell) = \exp\{-\text{Sim}(R_k^\ell, R_{k'}^\ell)\}$$

- Discourage similar record values



Artist	Artist
Yonder Mountain Band	Yonder Mountain

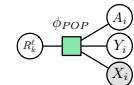


Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology

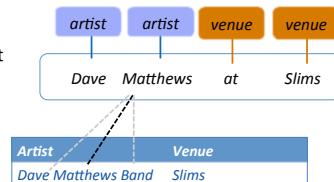


Term Popularity Factor



$$\phi_{POP}(x, y, R_A^\ell = v) = \sum_j \max_k \text{Sim}(x^j, y^j, v^k)$$

- Match each labeled message token to best record value token
- Token matching is IDF-weighted

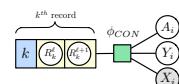


Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology

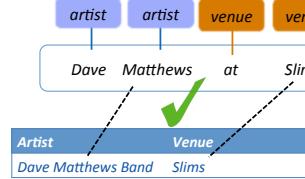


Record Consistency Factor



$$\phi_{CON}(x, y, R_A) = I[\phi_{POP}(x, y, R_A^\ell) > 0, \forall \ell]$$

- Encourage all record values to be in single message
- Active when there is some match for all record fields



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Experiments: Dataset

Twitter data: Three weekends of filtered messages:

- Authors from New York,
- Concert related messages (MIRA based classifier)

Resulting dataset: 5,800 messages

- Training – 2,184 messages (one weekend)
- Test – 3,662 messages (two weekends)

Gold event records:

- New York city events from NYC.com
- 11 events in training, 31 events in test.



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Experiment: Baselines

Baseline IE predictors.

- **List baseline:** String overlap with given list of artists and venues (Wikipedia)
- **CRF Voting baseline:** Extract record for each labeled pair of fields

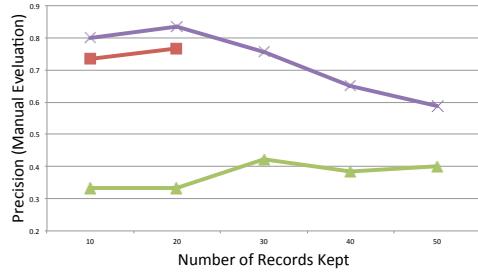


Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



Precision



Tackling the Challenges of Big Data

© 2014 Massachusetts Institute of Technology



