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
Information Extraction from Social Media

Regina Barzilay

Professor

Massachusetts Institute of Technology

My Research

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

Need in Information Extraction

- Lots of useful data in natural language

4 of 4 people found the following review helpful

★★★★★ **Fits perfect!**

By A. Graff 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

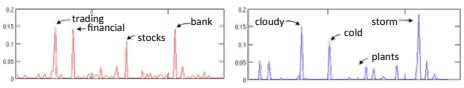
Information Extraction

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

- 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

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

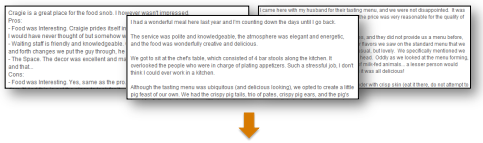
Information Extraction for Big Data

- 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

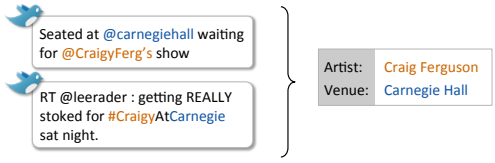
- Opportunity: Use Big Data to augment supervised learning
  - Uncover structured regularities in Big Data, which are relevant for the IE task

Multi-Aspect Summarization



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

Information Extraction from Tweets



Tackling the Challenges of Big Data
Big Data Analytics
Information Summarization
Information Extraction from Social Media

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Big Data Analytics

Regina Barzialy

Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

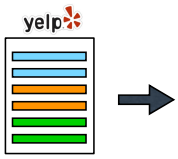
Tackling the Challenges of Big Data
Big Data Analytics
Information Summarization
Multi-aspect Summarization

Regina Barzilay

Professor

Massachusetts Institute of Technology

Multi-Aspect Summarization

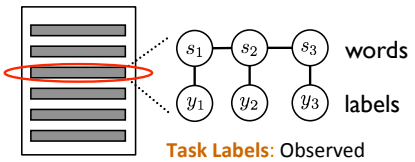


Aspect	Snippets
atmosphere	"stylish decor" "awesome art"
food	"loved it!" "tasty calzones!"
service	"fast and friendly" "impatient waiters"

Importance of Context:

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

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].

Adding Context

Sequence Labeling Task

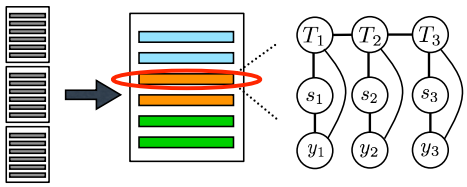
I ordered lunch from them the other day and I was [FOOD **pleasantly surprised**]. Our waiter dazzled me with his blue eyes and genuine smile, and all the waiters were [SERVICE **extremely professional and efficient**].

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.

Our Approach

Content Labels: Latent

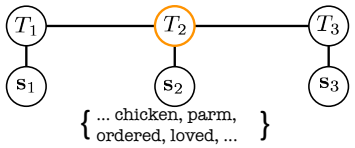


Task Labels: Observed

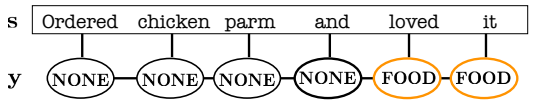
Goal: Use global content to assist with local decisions

Multi-Aspect Summarization

Content Model: Sentence-Level HMM



Task: Token-Level conditional random field



Factorization

$$P(\mathbf{T},\mathbf{s},\mathbf{y}) = P_\theta(\mathbf{T},\mathbf{s})P_\phi(\mathbf{y}|\mathbf{T},\mathbf{s})$$
$$= \prod_{i=1}^{n} P_\theta(T_{i+1}|T_i)\,(P_\theta(s_i|T_i)P_\phi(y_i|s_i,T_i))$$

Product over sentences — Topic Trans. — Bag-of-words — CRF

θ Topic Params
ϕ Task Params
$\mathbf{T}$ Sent. Topics
$\mathbf{y}$ Task Labels
$\mathbf{s}$ Sentences

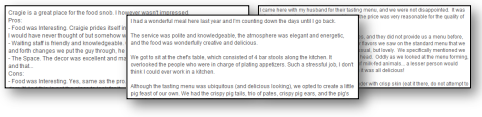
Data Set

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

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

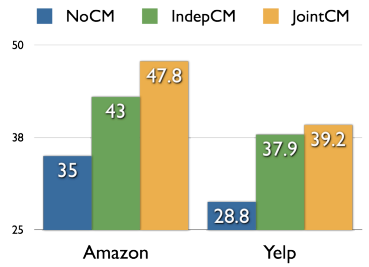
FOOD SERVICE ATMOSPHERE PRICE OVERALL	I came here with my husband for the tasting menu, and **we were not disappointed**. We got to sit at the chef's table, which overlooked the kitchen. The **service was polite and knowledgeable**, the **atmosphere was elegant and energetic** and the **food was wonderfully creative and delicious**.

Systems

- **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

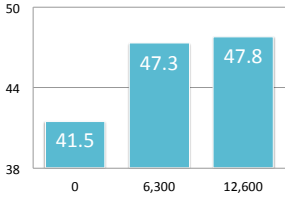
Results

Token F-measure Evaluation



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
Big Data Analytics
Information Summarization
Multi-Aspect Summarization

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Big Data Analytics

Regina Barzilay

Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Big Data Analytics
Information Summarization
Event Discovery in Social Media Feeds

Regina Barzilay

Professor

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*

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

IE for Social Media: Opportunity

Significant redundancy in Twitter stream:

Seated at @carnegiehall waiting for @CraigyFerg's show

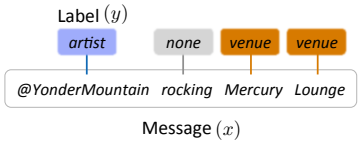
@DJPaulyD absolutely killed it at Terminal 5 last night.

Craig, nice seeing you #noelnight this weekend @becksdavis!

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

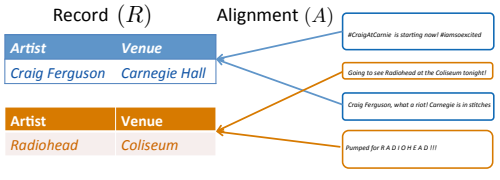
Model Functionality

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



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.



Model Overview

Source of supervision: Example event records
- Alignment between records and messages not observed.
- Message level field annotations not observed.

Free Music Fridays
July 16, 5:30pm at American Folk Art Museum

Glee Live
Jun 17, 8:00 PM at Izod Center

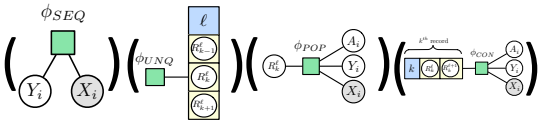
Indigo Girls
Jun 17, 8:00 PM at Tarrytown Music Hall

Model Overview

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

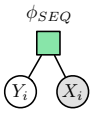
Learn jointly in factor graph model

$$P(R, A, y|x) \propto$$

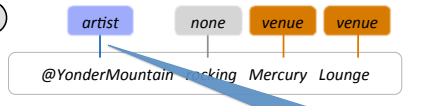


Sequence Labeling | Record Uniqueness | Term Popularity | Record Consistency

Sequence Labeling Factor



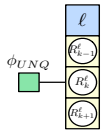
$$\phi_{SEQ}(x, y) = \exp\{\theta_{SEQ}^T f_{SEQ}(x, y)\}$$



- Similar to chain CRF
- Features on token and label
  - Wikipedia match, context, etc.

IsWikipediaMatch
word+1="rocking"
IsUserMention
....

Record Uniqueness Factor



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

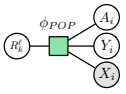
$$\phi_{UNQ}(R_k^\ell, R_{k'}^\ell) = \exp\{-\texttt{Sim}(R_k^\ell, R_k'^\ell)\}$$

- Discourage similar record values

Artist
Yonder Mountain Band

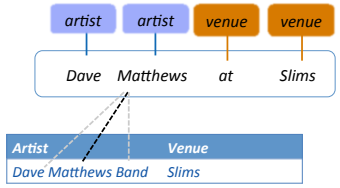
Artist
Yonder Mountain

Term Popularity Factor



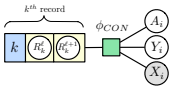
$$\phi_{POP}(x, y, R_A^\ell = v) = \sum_j \max_k \texttt{Sim}(x^j, y^j, v^k)$$

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



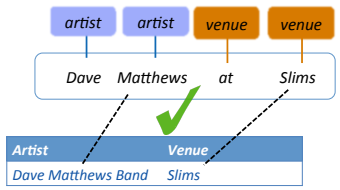
Artist	Venue
Dave Matthews Band	Slims

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



Artist	Venue
Dave Matthews Band	Slims

Experiments: Dataset

Twitter data: Three weekends of filtered messages:
- Authors from New York,
- Concert related messages (MIRA based classifier)

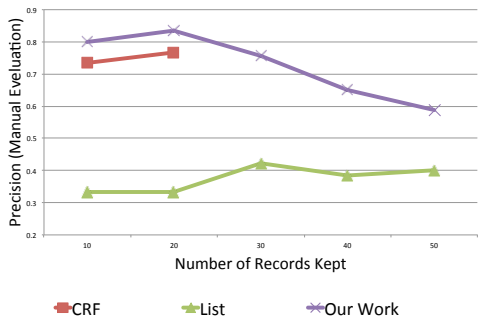
Resulting dataset: 5,800 messages
- Training – 2,184 messages (one weekend)
- Test – 3,662 messages (two weekends)

Gold event records:
- New York city events from NYC.com
- 11 events in training, 31 events in test.

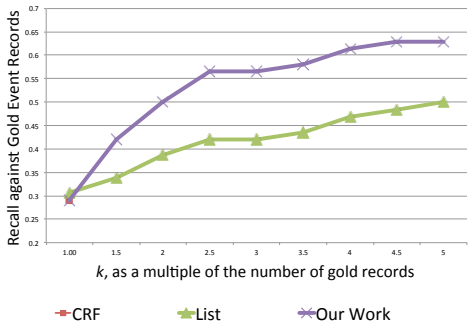
Experiment: Baselines

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

Precision



Recall



Conclusion

- Social media presents unique challenges and opportunities for NLP technologies
- Linguistically-rich models can compensate for noise inherent in social media streams
- Joint modeling of rich linguistic relations boosts prediction accuracy

Tackling the Challenges of Big Data
Big Data Analytics
Information Summarization
Event Discovery in Social Media Feeds

THANK YOU

PROFESSIONAL EDUCATION

CSAIL