Incorporating Social Context and Domain Knowledge for Entity Recognition

Jie Tang, Zhanpeng Fang
Department of Computer Science, Tsinghua University
Jimeng Sun
College of Computing, Georgia Institute of Technology
Entity Recognition in Social Media

- People use blogs, forums, and review sites to share opinions on politicians or products.
- One fundamental analytic issue is to recognize entity instances from the UGC short documents. However, the problem is very challenging
  - “S4” vs. “Samsung Galaxy S4”
  - “Fruit company” vs. “Apple Inc.”
  - “Peace West King” vs. “Xilai Bo” (a sensitive Chinese politician)
  - …
A Concrete Example

Challenges: short text + social networks + domain knowledge = ?
Related Work

• **Entity recognition**
  – Modeling as a ranking problem based on boosting and voted perceptron (Collins [9])
  – Incorporating long-distance dependency (Finkel et al. [13])
  – Use Labeled LDA [26] to exploit Freebase to help extraction (Ritter et al. [27])
  – Entity morph (Huang et al. [17])

• **Entity resolution**
  – A collective method for entity resolution in relational data (Bhattacharya and Getoor [4])
  – A hierarchical topic model for resolving name ambiguity (Kataria et al. [18])
  – Name disambiguation in digital libraries (Tang et al. [32])
Approach Framework
—SOCINST
Preliminary: Sequential Labeling

The input text $\mathbf{x}$

The label results $\mathbf{y}$

$\mathbf{y}^* = \max_y p(\mathbf{y} | \mathbf{x}; f, \Theta)$

where $f$ represents features and $\Theta$ are model parameters.
Sequential Labeling with CRFs

\[ p(y | x, \lambda, \mu) = \frac{1}{Z} \exp\left( \sum_i \sum_k \lambda_k f_k(x_i, y_i) + \sum_i \sum_j \mu_j f_j(x, y_i, y_{i+1}) \right) \]

\( \mu \) and \( \lambda \) are parameters to be learned from the training data.

\( f_k \) denotes the \( k \)-th feature defined for token \( x_i \)

\( f_j \) denotes the \( j \)-th feature defined for two consecutive tokens \( x_i \) and \( x_{i+1} \).
Sequential Labeling with CRFs

\[ \mathbb{P}(y | x, \lambda, \mu) = \frac{1}{Z} \exp(\lambda^k f_k(x_i, y_i) + \sum_{i} \sum_{j} \mu^j f_j(x_i, y_i, y_i+1)) \]

\[ f_k \text{ denotes the } k\text{-th feature defined for token } x_i \]
\[ f_j \text{ denotes the } j\text{-th feature defined for two consecutive tokens } x_j; \text{ and } x_j; \]

Performance of the model will be bad when dealing with short-text due to sparsity
Sequential Labeling Incorporating Topics

\[
p(y | x, \theta, \lambda, \mu) = \frac{1}{Z} \exp\left( \sum_i \sum_k \lambda_k f_k (x_i, \theta_i, y_i) + \sum_i \sum_j \mu_j f_j (x, \theta, y_i, y_{i+1}) \right)
\]
Latent Dirichlet Allocation

\[ p(x, z, \theta, \phi \mid \alpha, \beta) = \prod_{z=1}^{K} p(\phi_z \mid \beta) \prod_{d=1}^{M} p(\theta_d \mid \alpha) \prod_{i=1}^{N_d} p(x_i \mid \phi_z) p(z \mid \theta_d) \]

Extend to Model Authorship and Categories

• Generative process

Liberia Declared Free of Ebola
Shafiei and Milios

After the West African nation goes more than a month with no new reported cases of viral infection, the World Health Organization says the country is Ebola-free.

---

ACT Model

Generative process:

1. For each topic \( z \), draw \( \phi_z \) and \( \psi_z \) respectively from Dirichlet priors \( \beta_z \) and \( \mu_z \);

2. For each word \( w_{di} \) in document \( d \):
   - draw an author \( x_{di} \) from \( a_d \) uniformly;
   - draw a topic \( z_{di} \) from a multinomial distribution \( \theta_{x_{di}} \) specific to author \( x_{di} \), where \( \theta \) is generated from a Dirichlet prior \( \alpha \);
   - draw a word \( w_{di} \) from multinomial \( \phi_{z_{di}} \);
   - draw a category tag \( c_{di} \) from multinomial \( \psi_{z_{di}} \).

\[
P(z_{di}, x_{di} | x_{-di}, w, c, \alpha, \beta, \mu) \propto \\
\frac{m_{x_{di}} - d_{z_{di}} + \alpha_{z_{di}}}{\sum_z (m_{x_{di}} + \alpha_z)} \cdot \frac{n_{z_{di}w_{di}} - d_w + \beta_{w_{di}}}{\sum_v (n_{z_{di}w_{di}} + \beta_v)} \cdot \frac{n_{z_{di}c_{di}} - d_c + \mu_{c_{di}}}{\sum_c (n_{z_{di}c_{di}} + \mu_c)}
\]

Still challenges

However, we still cannot model domain knowledge and social context!

SOCINST: Modeling Domain Knowledge and Social Context Simultaneously
Modeling Domain Knowledge

\[
\alpha \rightarrow \theta_x \rightarrow \beta \rightarrow \phi_k \quad \text{for } k \in [1, K]
\]

\[
a_m \rightarrow x_{m,n} \rightarrow z_{m,n} \rightarrow c_{m,n} \quad \text{for } n \in [1, N_m], \quad m \in [1, M]
\]

\[
\text{DirichletTree}(\beta, \eta) = \left( \prod_{i=1}^{W} \phi_{z w_i}^{\eta w_i} \right) \times \\
\left( \prod_{j, c_j \in C} \frac{\Gamma(\sum_{k, w_k \in W(c)} \eta w_k)}{\prod_{k, w_k \in W(c)} \Gamma(\eta w_k)} \left( \sum_{k, w_k \in W(c)} \phi_{z w_i}^{\eta w_i} \right)^{\Delta(s)} \right)
\]

Modeling Social Context

$\alpha \xrightarrow{\theta_x} \beta \xrightarrow{\phi_k} \sum_{n \in [1, N_m]} x_{m,n} \xrightarrow{\gamma_j} \theta_{j\in NB(v_i)} \theta_{j}$

User A’s Social context is defined as a mixture of topic distributions of neighbors, i.e.

$$\sum_{j \in NB(v_i)} \gamma_j \theta_j$$
Theoretical Basis

• **Aggregation property** of Dirichlet distribution

If

\[(\theta_1, \ldots, \theta_i, \theta_{i+1}, \ldots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \alpha_K)\]

then

\[(\theta_1, \ldots, \theta_i + \theta_{i+1}, \ldots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_K)\]

• **Inverse of the aggregation property**

If

\[(\theta_1, \ldots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)\]

then

\[(\theta_1, \ldots, \tau \theta_i, (1-\tau)\theta_i, \ldots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \tau \alpha_i, (1-\tau)\alpha_i, \ldots, \alpha_K)\]
Model Learning

Input: a social network $G$, a document set $D$, a knowledge base $KB$;
Output: estimated parameters $\theta, \phi$

For each author $v$, draw $\theta_v$ from Dirichlet prior $\alpha$;
For each topic $z$, draw $\phi_z$ from Dirichlet prior $\beta$;

foreach document $d$ do
  if $v_d$ does not have relationship with others then
    foreach word $w_{d_i} \in w_d$ do
      Draw a topic $z_{d_i} \sim \text{multi}(\theta_v)$ from the topic model of user $v$;
      Call SamplingWord($z_{d_i}, w_{d_i}$);
    end
  end

else if $v_d$ have relationship with $v'$ then
  Construct a multinomial mixture $\theta_{v_d,v'}$ by combining topics distributions specific to users $v_d$ and $v'$;
  foreach word $w_{d_i} \in w_d$ do
    Draw a topic $z_{d_i} \sim \text{multi}(\theta)$ from the distribution specific to the pair;
    Call SamplingWord($z_{d_i}, w_{d_i}$);
  end
end

SamplingWord($z_{d_i}, w_{d_i}$)

if $w_{d_i}$ is an instance of a concept $c \in KB$ then
  Draw a concept path $\{c_k\}_k \sim \text{multi}(\pi)$ from a topic-specific concept path distribution;
  Draw word $w_{d_i} \sim \text{multi}(\psi_c)$ from a concept-specific multinomial distribution;
end
else
  Draw word $w_{d_i} \sim \text{multi}(\phi_{z_{d_i}})$ directly from a topic-specific multinomial distribution;
end

$$P(z_{d_i}|z_{-d_i}, w, \cdot) = \frac{n_v^{d_i} + \gamma n_v' z_{d_i} + \alpha}{\sum_z (n_v^{z_{d_i}} + \gamma n_v' z_{d_i}) + W\alpha} \times \prod_{k=1}^{T} \frac{m^{c_k}_{z_{di}c_{di}} + W_{c_k}^\beta}{\sum_{s} (m^{c_k}_{z_{di}c_{di}} + W_{c_k}^\beta)} \times \frac{m^{d_i}_{c_{di}w_{d_i}} + \eta}{\sum_{w} m^{d_i}_{c_{di}w_{d_i}} + W_c \eta}$$

$$P(z_{d_i}|z_{-d_i}, w, \cdot) = \frac{n_v^{d_i} + \alpha}{\sum_z n_v^{d_i} + W\alpha} \times \prod_{k=1}^{T} \frac{m^{c_k}_{z_{di}c_{di}} + W_{c_k}^\beta}{\sum_{s} (m^{c_k}_{z_{di}c_{di}} + W_{c_k}^\beta)} \times \frac{m^{d_i}_{c_{di}w_{d_i}} + \eta}{\sum_{w} m^{d_i}_{c_{di}w_{d_i}} + W_c \eta}$$

$$P(z_{d_i}|z_{-d_i}, w, \cdot) = \frac{n_v^{d_i} + \alpha}{\sum_z n_v^{d_i} + K\alpha} \times \frac{m^{d_i}_{z_{di}w_{d_i}} + \beta}{\sum_{w} m^{d_i}_{z_{di}w_{d_i}} + W\beta}$$
Sequential Labeling Incorporating Topics

\[ p(y | x, \theta, \lambda, \mu) = \frac{1}{Z} \exp(\sum_i \sum_k \lambda_k f_k(x_i, \theta_i, y_i) + \sum_i \sum_j \mu_j f_j(x, \theta, y_i, y_{i+1})) \]
Experiments
Data Sets

• All codes and datasets can be downloaded here http://aminer.org/socinst/

• Dataset

<table>
<thead>
<tr>
<th>Domain</th>
<th>#documents</th>
<th>#instances</th>
<th>#relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibo</td>
<td>1,800</td>
<td>545</td>
<td>10,763</td>
</tr>
<tr>
<td>I2B2</td>
<td>899</td>
<td>2,400</td>
<td>27,175</td>
</tr>
<tr>
<td>ICDM’12 Contest</td>
<td>2,110</td>
<td>565</td>
<td>NA</td>
</tr>
</tbody>
</table>

• Goal:
  – **Weibo**: Our goal is to extract real morph instances in the dataset.
  – **I2B2**: Our goal here is to extract private health information instances in the dataset.
  – **ICDM’12 Contest**: Our goal is to recognize product mentions in the dataset.
HISTORY OF PRESENT ILLNESS:
Mr. Blind is a 79-year-old white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum on November 13th at Sephsandpot Center. The patient developed hematemesis on November 15th and was intubated for respiratory distress. He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm.
ICDM’12 Contest

Fourm

A TV dealer in the area reports that some Philips HD sets can't decode a picture on 15-1...

Hi I wondering if this new Pioneer receiver, SC-57 is as powerful as SC-37 was/is?

but when I tried it in the PS3 it came up "no files". Then after I read DoItYourself's suggestion...

The HD65 is a good machine but it cannot match the quality of the IN80 or BenQ W5000.

SubTask1: Identify the product mentions

SubTask2: Align mentions to product entries

Catalogs

- Philips 46pfl4706 46" 1080p HD Led Lcd Television
- DecalGirl PS3S-MICROBIA PS3 Slim Skin - Microbia
- Sony Playstation 3 250GB - PS3 Console
- Philips 40pfl7705dv 40" Hd Lcd Television
- Optoma Hd65 Dip Projector
- Philips 40pfl5706 F7 40-inch 1080p Lcd Tv With Pixel Precise Hd
- Playstation Ps3 3d Display Bundle W/ (5) 3d Games Mlb 12 + (2) 3d Glasses
- Optoma Hd65 Dip Projector - Working But Fair Condition
- Harmony Adapter for PS3
Results

- **SM**: Simply extracts all the terms/symbols that are annotated
- **RT**: Recognizes target instances from the test data by a set of rule templates
- **CRF**: Trains a CRF model using features associated with each token
- **CRF+AT**: Uses Author-Topic (AT) [30] to train a model and then it use the learned topics as features for CRF for instance recognition
- **SOCINST**: Our proposed model
## Results

- **SM**: Simply extracts all the terms/symbols that are annotated.
- **RT**: Recognizes target instances from the test data by a set of rule templates.
- **CRF**: Trains a CRF model using features associated with each token.
- **CRF+AT**: Uses Author-Topic (AT) [30] to train a model and then it use the learned topics as features for CRF for instance recognition.
- **SOCINST**: Our proposed model.

<table>
<thead>
<tr>
<th>Data</th>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Weibo</strong></td>
<td>SM</td>
<td>55.34</td>
<td>34.92</td>
<td>42.82</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>39.62</td>
<td>66.31</td>
<td>49.60</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>29.24</td>
<td>94.89</td>
<td>44.71</td>
</tr>
<tr>
<td></td>
<td>CRF+AT</td>
<td>43.71</td>
<td>89.67</td>
<td>58.77</td>
</tr>
<tr>
<td></td>
<td>SOCINST</td>
<td><strong>65.72</strong></td>
<td>76.27</td>
<td><strong>70.60</strong></td>
</tr>
<tr>
<td><strong>I2B2</strong></td>
<td>SM</td>
<td>39.58</td>
<td>28.24</td>
<td>32.96</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>39.60</td>
<td>40.29</td>
<td>39.94</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>40.99</td>
<td>56.19</td>
<td>47.40</td>
</tr>
<tr>
<td></td>
<td>CRF+AT</td>
<td>41.37</td>
<td>54.92</td>
<td>47.19</td>
</tr>
<tr>
<td></td>
<td>SOCINST</td>
<td><strong>43.94</strong></td>
<td><strong>57.18</strong></td>
<td><strong>49.69</strong></td>
</tr>
<tr>
<td><strong>ICDM’12 Contest</strong></td>
<td>SM</td>
<td>9.47</td>
<td>62.50</td>
<td>16.46</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>23.69</td>
<td>42.01</td>
<td>30.30</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>21.80</td>
<td>53.48</td>
<td>30.97</td>
</tr>
<tr>
<td></td>
<td>CRF+AT</td>
<td>26.54</td>
<td>51.37</td>
<td>35.00</td>
</tr>
<tr>
<td></td>
<td>SOCINST</td>
<td><strong>37.91</strong></td>
<td><strong>53.33</strong></td>
<td><strong>44.32</strong></td>
</tr>
</tbody>
</table>
More Results—ICDM’12 Contest

Performance comparison of SOCINST and the first place [38] in ICDM’12 Contest.

By incorporating the modeling results into the CRF model

Effects of Social Context and Domain Knowledge

$\text{SOCINST}_{\text{base}}$ — we removed both social context and domain knowledge from our method;

$\text{SOCINST-SC}$ — we removed social context from our method;

$\text{SOCINST-DK}$ — we removed domain knowledge from our method;
Parameter Analysis

- **Graph 1**: F1-Measure vs. Number of topics
  - The F1-Measure increases with the number of topics initially, reaching a peak at around 14 topics, and then decreases slightly.

- **Graph 2**: F1-Measure vs. Number of iterations
  - The F1-Measure increases rapidly from 50 to 100 iterations, then fluctuates between 100 and 200 iterations, and finally stabilizes around 250 iterations.
Parameter Analysis (cont.)

* All the other hyperparameters fixed
The number of topics is set to $K = 15$
Chapter 1: Introduction to Data Mining

We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored in disparate structures very rapidly became overwhelming. This initial data has led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the “essence” of information stored, and the discovery of patterns in raw data.

What kind of information are we collecting?

We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypermedia documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files.

- Business transactions: Every transaction in the business industry is often “monitored” for profitability. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra-business operations such as management of in-house warehouses and assets. Large department stores, for example, thanks to the widespread use of bar codes, store millions of transactions daily representing often iterations of data. Storage space is not the major problem, as the price of hard disks is continuously dropping, but the effective use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world.

- Scientific data: Whether in a Swiss nuclear accelerator laboratory counting particles in the Canadian forest underbeta readings from a extinct bear radio collar,
Conclusion

• Study the problem of instance recognition by incorporating social context and domain knowledge

• Propose a topic modeling approach to learn topics by considering social relationships between users and context information from a domain knowledge base

• Experimental results on three different datasets validate the effectiveness and the efficiency of the proposed method.
Future work

• The general idea of incorporating social context and domain knowledge for entity recognition represents a new research direction
• Combining the sequential labeling model and the proposed SOCINST into a unified model should be beneficial
• Further incorporating other social interactions, such as social influence, to help instance recognition is an intriguing direction
Thank you!

Collaborators:
Jimeng Sun (Georgia Tech)
Zhanpeng Fang (THU)

Jie Tang, KEG, Tsinghua U,
Download all data & Codes,
http://keg.cs.tsinghua.edu.cn/jietang
http://aminer.org/socinst
Modeling Short Text with Topics

\[ p_d(x) = \lambda_B p(x | \theta_B) + (1 - \lambda) \sum_{k=1}^{K} \pi_{d,k} p(x | \theta_k) \]

\[ \log p(d) = \sum_{x \in V} n(x,d) \log[\lambda_B p(x | \theta_B) + (1 - \lambda) \sum_{k=1}^{K} \pi_{d,k} p(x | \theta_k)] \]

"Generating" word x in doc d in the collection

Parameters:
\[ \theta_B = \text{noise-level (manually set)} \]
\[ \theta_1 \text{ and } \pi \text{ are estimated with Maximum Likelihood} \]
\[ a \rightarrow \theta_x \rightarrow \beta \rightarrow \phi_k \land k \in [1, K] \]

\[ a_m \rightarrow x_{m,n} \rightarrow z_{m,n} \rightarrow x_{m,n} \rightarrow c_{m,n} \land n \in [1, N_m] \]

\[ m \in [1, M] \]