Named Entity Recognition and Linking

Dr. SUN Aixin  孙爱欣
School of Computer Science and Engineering
NTU, Singapore
NER and EL

- Named-entity recognition (NER)
  - The task to locate and classify named entities in text into pre-defined categories
    - names of persons, organizations, locations,
    - expressions of times, quantities, monetary values, percentages, etc.
  - Example: [Jim]_{Person} bought 300 shares of [Acme Corp.]_{Organization} in [2006]_{Time}.

- Entity linking (EL)
  - The task of determining the identity of entities mentioned in text, with reference to a knowledge base.
  - Example: Michael Jordan will give a talk at the conference
NER from Text

• Formal text (news papers, research articles)
  – Lexical features
  – Grammatical features
  – …

• Social media
  – Informal language
  – Misspellings
  – Grammatical errors
  – Self-defined abbreviations
  – And many others….
NER from Social Media

• Domain-specific knowledge in **User Language**
  – Collection of terms used by users to name entities in a specific domain
  – Domain → defines term meanings

• Why not general (open-domain) knowledge bases?
  – Wikipedia, Freebase, ProBase …
  – What does this term mean: “**TCU 2/52**”

• Case study:
  – Extract mobile phone names from user forum
  – Location extraction from tweets
### Samsung Galaxy SIII – real data from Singaporean users

<table>
<thead>
<tr>
<th>Name variation</th>
<th>#users</th>
</tr>
</thead>
<tbody>
<tr>
<td>galaxy s3</td>
<td>553</td>
</tr>
<tr>
<td>s3 lte</td>
<td>343</td>
</tr>
<tr>
<td>samsung galaxy s3</td>
<td>284</td>
</tr>
<tr>
<td>s iii</td>
<td>242</td>
</tr>
<tr>
<td>galaxy s iii</td>
<td>225</td>
</tr>
<tr>
<td>samsung s3</td>
<td>219</td>
</tr>
<tr>
<td>sgs3</td>
<td>187</td>
</tr>
<tr>
<td>siii</td>
<td>149</td>
</tr>
<tr>
<td>samsung galaxy s iii</td>
<td>145</td>
</tr>
<tr>
<td>i9300</td>
<td>120</td>
</tr>
<tr>
<td>gs3</td>
<td>82</td>
</tr>
<tr>
<td>galaxy siii</td>
<td>61</td>
</tr>
<tr>
<td>i9305</td>
<td>52</td>
</tr>
<tr>
<td>lte s3</td>
<td>46</td>
</tr>
<tr>
<td>galaxy s3 lte</td>
<td>45</td>
</tr>
<tr>
<td>s3 non lte</td>
<td>32</td>
</tr>
<tr>
<td>samsung galaxy s iii</td>
<td>32</td>
</tr>
<tr>
<td>sgs 3</td>
<td>27</td>
</tr>
<tr>
<td>samsung galaxy s3 lte</td>
<td>22</td>
</tr>
<tr>
<td>sg3</td>
<td>21</td>
</tr>
<tr>
<td>gsiii</td>
<td>16</td>
</tr>
<tr>
<td>samsung galaxy s3 i9300</td>
<td>15</td>
</tr>
<tr>
<td>samsung i9300 galaxy s iii</td>
<td>13</td>
</tr>
<tr>
<td>s3 4g</td>
<td>11</td>
</tr>
<tr>
<td>3g s3</td>
<td>11</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Words grouping by Brown Clustering

<table>
<thead>
<tr>
<th>Code</th>
<th>Example Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>^0110111001 (43)</td>
<td>everything everythin everthing evrything everythang everythinq everythng everythng everythng everythngg everythng errthang errything #everything errthing erything everythnggg errthang evrythin erthang 1thing erthing everythnggg jony everythig everytin everythng everyithng everythign everythn everyhting everythinqq everythink erythang everythng_everythngggg errrthang everythg everyword eyrethinyg everysong er'thing</td>
</tr>
<tr>
<td>^0110111010 (85)</td>
<td>nothing nothin nun nuthin nuttin nuffin 10x noting nthn nowt nuthn nuthing 100x nothingg nothn nothinq nutin nutin nuffin nutn nutn whatever #nodisrespect nuttn nothinggg #dontmeantobrag 1000x zilch nothinn #nothing nothng nutting nufin nuin nout nothinqq nthng nthing nothingggg nufin nofin nothen nthtn nothinr ntn nought ntg nothinnn nothign n0thing nothig</td>
</tr>
<tr>
<td>^01101110101 (108)</td>
<td>something somethin sumthin sumthing sumn somthing sth sumtn sumtin smth somthin suttin sumnin somethinq summin treatser somethingg someting sumfin smthn somethn somethng summat smthng smthing sum'n sumtnhng sum'g somn sumtn smething sometin somethign sum10 something somethinn somethinggg somethig sumpin somin sumting something/someone somtin somehting somthn someshit sumptin something- smtg thereabouts</td>
</tr>
<tr>
<td>^01101110111 (36)</td>
<td>anything anythin nething anythng anythingg anythingg anythingg anythingg anythingg anythingg anfian vaart anything anythign nethin nything anyhting #anything anything- anythingg anythinggg nothing- anything anythink anything anythingqq somethingggg nthng anything woods enendsmeat anything/anyone anything anything anything enything</td>
</tr>
</tbody>
</table>

Source: [http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html](http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html)
Dictionary: candidate mobile phone names

Brown Clustering

Sentences in forum threads

Sentence Parser

Brand word clusters containing “samsung”, “htc”, ...

<table>
<thead>
<tr>
<th>229</th>
<th>hitec</th>
<th>rovio</th>
<th>samsun</th>
<th>ssg</th>
</tr>
</thead>
<tbody>
<tr>
<td>503</td>
<td>klipsch</td>
<td>sam</td>
<td><strong>samsung</strong></td>
<td>sumsung</td>
</tr>
<tr>
<td>andino</td>
<td>magpul</td>
<td>sammy</td>
<td>sumsung</td>
<td>sung</td>
</tr>
<tr>
<td>aston</td>
<td>msgtypes</td>
<td>samseng</td>
<td>samung</td>
<td>vivendi</td>
</tr>
<tr>
<td>cnc</td>
<td>msung</td>
<td>samsuck</td>
<td>seagate</td>
<td>wright</td>
</tr>
<tr>
<td>fujifilm</td>
<td>netgear</td>
<td>samsung</td>
<td>semaphore</td>
<td></td>
</tr>
</tbody>
</table>

Model word clusters containing “galaxy”, “active” ...

Noun phrases

- samsung galaxy s3
- samsung firmware
- service center
- Samsung updates...

Candidate Name Filter (2 Rules)

Candidate names

- samsung galaxy s3, gs3
- sgs3, s3 lte
- samsung updates
- service center ...

Brand Filter (4 Rules)

Brand variations

- ssg, samseng, sam, **samsung**, sammy, sumsung, samsun, sung, samsuck, samsung, sumsungs, samung
### Dictionary (knowledge) in user language

<table>
<thead>
<tr>
<th>Brand</th>
<th>User spellings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple, HTC, LG</td>
<td>–No brand variations–</td>
</tr>
<tr>
<td>Nokia</td>
<td>nokia, nokie, nk</td>
</tr>
<tr>
<td>BlackBerry</td>
<td>blackberry, bbry, blackberry, bb, bberry</td>
</tr>
<tr>
<td>Samsung</td>
<td>ssg, samseng, sam, samsung, sammy, sumsung, samsun, sung, samsuck, samsungs,</td>
</tr>
<tr>
<td></td>
<td>samung</td>
</tr>
<tr>
<td>Sony Ericsson</td>
<td>sony ericson, sony ericsson, sony ericson, sony ericcson, sonyericsson, sony</td>
</tr>
<tr>
<td></td>
<td>ericsson, sn, sony, sonyeric</td>
</tr>
<tr>
<td>Motorola</td>
<td>motorola, moto, motorolla, mot</td>
</tr>
</tbody>
</table>

- Many variants
- Many users do use formal names
- Brand, series, model
- The usage context shall be similar
Recognize names based on a dictionary in user language

- Generate **candidate names** based on naming convention
- Recognize true product names from candidate names
- Normalize names based on naming convention

Diagram:

1. Candidate Name Generator
2. CRF-based Name Recognizer
3. Rule-based Name Normalizer

Output:

- Mobile Phone Formal Names

Keywords:

- Samsung
- Apple
- Microsoft
- Nokia
- Sony
- LG
- HTC
- Motorola
- Huawei

Latest Devices:

- Meizu m5 Note
- HTC Desire 650
- Meizu Pro 7
- HTC 10 evo
- OnePlus 3T
- Apple iPhone 8
Named Entity Linking

Although the shots sounded the death-knell for the Pelicans, they were greeted by cheers from fans, who like their counterparts around the country, have rallied behind them during a season that has turned into a farewell to those adoring fans a glimpse of past glories.

"He's on a nice little roll," said Lakers coach Byron Scott. "Our young guys are still so young they don't understand when you've got a double-digit lead you can't relax," Scott said. "Not in this league."

Candidates (local confidence):
- New_Orleans_Pelicans (0.28)
- Lahti_Pelicans (0.07)
- Pelican (0.04)
- #Pelicans (0.02)
- Perth_Pelicans (0.01)
- New_Orleans_Pelicans_(baseball) (0.01)
- Australian_pelican (0.01)
- Myrtle_Beach_Pelicans (0.01)

The linking is made at step 35. Click for more details.
Named Entity Linking

- Local confidence vs collective context
Collective Linking

• Collective linking:
  – Utilize semantic relatedness to improve linking performance
  – e.g. “Wood played at 2006 Masters held in Augusta, Georgia”.

• Semantic relatedness measures
  – Jaccard Similarity (JS) of citing article sets
  – Entity Embedding Similarity (EES)
Collective Linking: Assumption

• All pairs of linked entities are related:

\[
\phi(m_i, e_i) \quad \psi(e_i, e_j)
\]

“Wood played at 2006 Masters held in Augusta, Georgia”

Complete-pairwise coherence model

\[
\Gamma^* = \arg \max_{\Gamma} \left[ \sum_{i=1}^{N} \phi(m_i, e_i) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \psi(e_i, e_j) \right]
\]

Local confidence \quad Global coherence
Are mentioned entities densely connected?

“Wood played at 2006 Masters held in Augusta, Georgia”
Are mentioned entities densely connected?

“The Sun and The Times reported that Greece will have to leave the Euro soon.”

Complete-pairwise coherence is not always necessary
Complete-pairwise coherence is not always necessary?

- Measure the **degree of coherence** in real datasets
  - Average degree of entity relatedness graph which consists of high-weighted edges (by JS or EES measures).

\[ 2 \frac{N - 1}{N} \]  
\[ \frac{N - 1}{2N - 1} \]

(a) Dense  
(b) Tree-like  
(c) Chain-like  
(d) Forest-like
In general, the calculated values lie closer to tree (or chain) form’s expected values rather than that of the dense form.
Our Idea: Pair-Linking

• We do not need to look at all other entity when deriving linking decisions.

• Interactively resolve a pair of mention at each step, from the more confident pairs to less confident pairs.

“Wood played at 2006 Masters held in Augusta, Georgia”
Pair-Linking: Local confidence + Coherence

• Pairwise confidence

\[ \phi(m_i, e_i) \]

\[ \psi(e_i, e_j) \]

\[ \phi(m_j, e_j) \]

Wood

USS Wood (ship)

Wood, Wisconsin (town)

Singapore_Masters

Masters of Horror (movie)

2006 Master
Pair-Linking Example

“Wood played at 2006 Masters held in Augusta, Georgia”

Wood 0.75 0.6 0.4
USS Wood (ship) 0.7 0.9
Tiger Wood

Augusta
USS Augusta
Augusta University
Augusta, Georgia

Georgia
Georgia (country)
University of Georgia
Georgia, U.S. State

2006 Masters Tournament
2006 Masters

Confidence \( m_i \rightarrow e_i \)
\( m_j \rightarrow e_j \)
Pair-Linking Example

“Wood played at 2006 Masters held in Augusta, Georgia”
“Wood played at 2006 Masters held in Augusta, Georgia”
“Wood played at 2006 Masters held in Augusta, Georgia”
“Wood played at 2006 Masters held in Augusta, Georgia”
Pair-Linking Example

“Wood played at 2006 Masters held in Augusta, Georgia”
Pair-Linking is Super Fast

• Pair-Linking cares about the pair with highest confidence score.
  – Use priority queue to store and retrieve the pair.
  – Utilize early stop to avoid scanning all possible pair of candidates.
## Experiment: 8 benchmark datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Type</th>
<th>#documents</th>
<th>Avg #words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuters128</td>
<td>News</td>
<td>111</td>
<td>136</td>
</tr>
<tr>
<td>ACE2004</td>
<td>News</td>
<td>35</td>
<td>375</td>
</tr>
<tr>
<td>MSNBC</td>
<td>News</td>
<td>20</td>
<td>544</td>
</tr>
<tr>
<td>DBpedia</td>
<td>News</td>
<td>57</td>
<td>29</td>
</tr>
<tr>
<td>RSS500</td>
<td>RSS-feeds</td>
<td>343</td>
<td>30</td>
</tr>
<tr>
<td>KORE50</td>
<td>Short sentence</td>
<td>50</td>
<td>12</td>
</tr>
<tr>
<td>Micro2017</td>
<td>Tweets</td>
<td>696</td>
<td>18</td>
</tr>
<tr>
<td>AQUAINT</td>
<td>News</td>
<td>50</td>
<td>220</td>
</tr>
</tbody>
</table>
Pair-Linking Performance

- Linking accuracy (F1)

<table>
<thead>
<tr>
<th>CL Method</th>
<th>Reuters128</th>
<th>ACE2004</th>
<th>MSNBC</th>
<th>Dbpedia</th>
<th>RSS500</th>
<th>KORE50</th>
<th>Micro2014</th>
<th>AQUAINT</th>
<th>Average</th>
<th>#1st</th>
<th>#2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iter_Sub(AL)</td>
<td>0.856</td>
<td>0.894</td>
<td>0.879</td>
<td>0.839</td>
<td>0.793</td>
<td>0.682</td>
<td>0.811</td>
<td>0.876</td>
<td>0.829</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Iter_Sub(SL)</td>
<td>0.807</td>
<td>0.883</td>
<td>0.870</td>
<td>0.835</td>
<td>0.809</td>
<td>0.653</td>
<td>0.808</td>
<td>0.850</td>
<td>0.814</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LBP(AL)</td>
<td><strong>0.864</strong></td>
<td>0.861</td>
<td>0.895</td>
<td>0.833</td>
<td>0.777</td>
<td>0.715</td>
<td>0.822</td>
<td>0.877</td>
<td>0.831</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>LBP(SL)</td>
<td>0.823</td>
<td>0.875</td>
<td>0.900</td>
<td>0.843</td>
<td>0.814</td>
<td>0.762</td>
<td><strong>0.824</strong></td>
<td>0.872</td>
<td>0.839</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>FwBw</td>
<td>0.830</td>
<td><strong>0.895</strong></td>
<td>0.905</td>
<td>0.832</td>
<td>0.802</td>
<td>0.749</td>
<td>0.818</td>
<td>0.866</td>
<td>0.837</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.837</td>
<td>0.882</td>
<td>0.888</td>
<td>0.822</td>
<td>0.785</td>
<td>0.512</td>
<td>0.797</td>
<td>0.872</td>
<td>0.799</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pair-Linking</td>
<td><strong>0.859</strong></td>
<td>0.883</td>
<td><strong>0.910</strong></td>
<td>0.845</td>
<td>0.823</td>
<td>0.787</td>
<td>0.813</td>
<td><strong>0.879</strong></td>
<td><strong>0.850</strong></td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

- Speed: time per document in millisecond

<table>
<thead>
<tr>
<th>CL method</th>
<th>Reuters128</th>
<th>ACE2004</th>
<th>MSNBC</th>
<th>Dbpedia</th>
<th>RSS500</th>
<th>KORE50</th>
<th>Micro2014</th>
<th>AQUAINT</th>
<th>#1st</th>
<th>#2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iter_Sub(AL)</td>
<td>97.515</td>
<td>21.369</td>
<td>3010.214</td>
<td>12.922</td>
<td>0.127</td>
<td>2.235</td>
<td>0.682</td>
<td>293.271</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Iter_Sub(SL)</td>
<td>67.772</td>
<td>20.183</td>
<td>3211.341</td>
<td>11.603</td>
<td>0.108</td>
<td>2.284</td>
<td>0.684</td>
<td>107.640</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LBP(AL)</td>
<td>40.049</td>
<td>41.911</td>
<td>1584.504</td>
<td>42.673</td>
<td>0.331</td>
<td>11.515</td>
<td>3.667</td>
<td>269.854</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LBP(SL)</td>
<td>92.625</td>
<td>43.173</td>
<td>4421.172</td>
<td>44.263</td>
<td>0.289</td>
<td>8.627</td>
<td>3.170</td>
<td>403.140</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FwBw</td>
<td><strong>0.940</strong></td>
<td>1.975</td>
<td><strong>8.880</strong></td>
<td>2.034</td>
<td>0.103</td>
<td>1.190</td>
<td>0.367</td>
<td>4.959</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>PageRank</td>
<td>110.572</td>
<td>77.398</td>
<td>4293.670</td>
<td>132.009</td>
<td>5.436</td>
<td>64.982</td>
<td>15.796</td>
<td>375.239</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pair-Linking</td>
<td><strong>1.721</strong></td>
<td><strong>0.590</strong></td>
<td><strong>28.699</strong></td>
<td><strong>0.491</strong></td>
<td><strong>0.025</strong></td>
<td><strong>0.951</strong></td>
<td><strong>0.117</strong></td>
<td><strong>3.105</strong></td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

(*) Performances on ACE2004, RSS500 and Micro2014 are not shown here.
• Contributors
  – Phan Cong Minh
  – Han Jialong
  – Tay Yi
  – Li Chenliang
  – Yao Yangjie

• *Mobile Phone Name Extraction from Internet Forums: A Semi-supervised Approach.*

• *NeuPL: Attention-based Semantic Matching and Pair-Linking for Entity Disambiguation*
  Minh C. Phan, Aixin Sun, Yi Tay, Jialong Han, Chenliang Li. CIKM 2017

• *Pair-Linking for Collective Entity Disambiguation: Two Could Be Better Than All*
  Minh C. Phan, Aixin Sun, Yi Tay, Jialong Han, Chenliang Li. ArXiv

• Project demo: [https://youtu.be/w3EsALNrKAk](https://youtu.be/w3EsALNrKAk)