Automatic Knowledge Base Construction Systems

Dr. Daisy Zhe Wang
CISE Department
University of Florida
September 3th 2014
Text Contains Knowledge
Text Contains Automatically Extractable Knowledge
Outline

• Knowledge Base construction from Text
  – Information Extraction (IE)
  – Natural language processing (NLP)
Knowledge/Information Extraction: Named Entity Recognition

• “We are pleased that today's agreement guarantees our corporation will maintain a significant and long term presence in the Big Apple," McGraw-Hill president Harold McGraw III said in a statement.

--- From New York Times April 24, 1997
Knowledge/Information Extraction: Named Entity Recognition

• “We are pleased that today's agreement guarantees our corporation will maintain a significant and long term presence in the Big Apple,” McGraw-Hill president Harold McGraw III said in a statement.

--- From New York Times April 24, 1997

Labels:
Person  Company  Location  Other
Common NLP/IE Tasks

- POS Tagging
- Shallow parsing/chunking
- NER (Named entity recognition)
- Co-reference (intra-, cross- document)
- Relation extraction
- Event Extraction
Sequence Labeling: The Problem

• Given a sequence (in NLP, words), assign appropriate labels to each word.

• Named-entity recognition (NER)

“We are pleased that today's agreement guarantees our corporation will maintain a significant and long term presence in the Big Apple," McGraw-Hill president Harold McGraw III said in a statement.

--- From New York Times April 24, 1997
Sequence Labeling: The Problem

• Given a sequence (in NLP, words), assign appropriate labels to each word.

• Part-of-speech POS tagging:

  DT NN VBD IN DT NN .
  The cat sat on the mat.

  – 36 part-of-speech tags used in the Penn Treebank Project:
    https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
Sequence Labeling: The Problem

• Given a sequence (in NLP, words), assign appropriate labels to each word.

• Shallow/partial parsing (aka chunking): which groups of words go together as "phrases"

B-NP I-NP B-VP B-PP B-NP I-NP
The cat sat on the mat
Sequence Labeling: The Problem

• Given a sequence (in NLP, words), assign appropriate labels to each word.

• Another example, relation extraction (e.g., arguments and relation types): which words are the subject or object of a verb

   B-Arg I-Arg B-Rel I-Rel B-Arg I-Arg
   The cat sat on the mat
   Madam Curie won Nobel Prize
Sequence Labeling Models

HMM

Generative

Conditional

Discriminative

Linear-chain CRF
A Graphical Model – Conditional Random Fields (CRF)

Text (address string):
E.g., “2181 Shattuck North Berkeley CA USA”

CRF Model:

X=tokens

Y=labels

Possible Extraction Worlds:

x  2181  Shattuck  North  Berkeley  CA  USA
y1  apt. num  street name  city  city  state  country  (0.6)
y2  apt. num  street name  street name  city  state  country  (0.1)
Viterbi Algorithm – MAP labeling

Viterbi Dynamic Programming Algorithm:

\[
V(i, y) = \begin{cases} 
\max_{y'} (V(i-1, y') + \sum_{k=1}^{K} \lambda_k \cdot f_k(y, y', x_i)), & \text{if } i \geq 0 \\
0, & \text{if } i = -1.
\end{cases}
\]

<table>
<thead>
<tr>
<th>pos</th>
<th>street num</th>
<th>street name</th>
<th>city</th>
<th>state</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>2181</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12</td>
<td>24</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>21</td>
<td>32</td>
<td>32</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>29</td>
<td>40</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>39</td>
<td>47</td>
<td>46</td>
<td>50</td>
</tr>
</tbody>
</table>
IE/NLP Libraries

• NLTK
• LingPipe
• Stanford Parser
• Apache OpenNLP
• Illinois NLP Software: http://cogcomp.cs.illinois.edu/page/software
• CRF project: http://crf.sourceforge.net/
• MADLib CRF: http://doc.madlib.net/v0.7/group__grp__crf.html
• GATE, TRIPS, ...
Knowledge Bases

- A *knowledge base* is a collection of entity, facts, relationships that conforms with a certain data model.
- A knowledge base helps machine understand humans, languages, and the world.
- Examples 1: Google Knowledge Graph
Knowledge Bases From Big Data


- KB Applications:
  - Improve Search Engine (e.g., Google, Bing)
  - Automatically populating Wikipedia (e.g., references, info box)
  - Domain-specific Knowledge Bases (e.g., UMLS)
TREC Knowledge Base Acceleration Pipeline (GatorDSR)

1. Streaming/Data Processing System
2. POS
3. Chunking
4. Named entity extraction
5. Co-reference
6. Relation extraction
7. Slot filling
Knowledge Bases (KBs)

- Freebase
- YAGO
- NELL
- TextRunner/ReVerb
- DBpedia
- ProBase
- WordNet/VerbNet
- Cyc, ConceptNet– common sense KBs
Querying Knowledge Bases

- SPARQL over JENA RDF store/ OpenLink Virtuoso