600.465 Connecting the dots - II (NLP in Practice)

Delip Rao
Master subtitle style
delip@jhu.edu

Last class ...

- Understood how to solve and ace in NLP tasks
- General methodology or approaches
- End-to-end development using an example task
  - Named Entity Recognition

Shared Tasks: NLP in practice

- Shared Task (aka Evaluations)
  - Everybody works on a (mostly) common dataset
  - Evaluation measures are defined
  - Participants get ranked on the evaluation measures
  - Advance the state of the art
  - Set benchmarks
- Tasks involve common hard problems or new interesting problems

Person Name Disambiguation

- Computational linguist
- Physicist
- Psychologist
- Architect
- Lawyer
- Sculptor
- Biologist
- Musician
- CEO
- Tennis player
- Theologist
- Pastor

Rao, Garera & Yarowsky, 2007

Clustering using web snippets

Goal: To cluster 100 given test documents for name “David Smith”

Step 1: Extract top 1000 snippets from Google

Step 2: Cluster the 1100 documents together

Step 3: Extract the clustering of the test documents

Rao, Garera & Yarowsky, 2007
Web Snippets for Disambiguation

- Snippets contain high quality, low noise features
- Easy to extract
- Derived from sources other than the document (e.g., link text)

Rao, Ganra & Yarowsky, 2007

Term bridging via Snippets

Snippets contain both the terms “780 492-9920” and “T6G2H1” and that can serve as a bridge for clustering Document 1 and Document 2 together

Rao, Ganra & Yarowsky, 2007

Challenges in Entity Linking

- Name Variation
  - Abbreviations: BSO vs. Boston Symphony Orchestra
  - Shortened forms: Osama Bin Laden vs. Bin Laden
  - Alternate spellings: Osama vs. Ussamah vs. Oussama
- Entity Ambiguity: Polysemous mentions
  - E.g., Springfield, Washington
- Absence: Open domain linking
  - Not all observed mentions have a corresponding entry in KB (NIL mentions)

Entity Linking

- Identify matching entry, or determine that entity is missing from KB

Entity Linking: Features

- Name-matching
  - acronyms, aliases, string-similarity, probabilistic FST
- Document Features
  - TF/IDF comparisons, occurrence of names or KB facts in the query text, Wikitology
- KB Node
  - Type (e.g., is this a person), Features of Wikipedia page, Google rank of corresponding Wikipedia page

Absence (NIL indications)
Entity Linking: Name Matching

- Acronyms
- Alias Lists
  - Wikipedia redirects, stock symbols, misc. aliases
- Exact Match
  - With and without normalized punctuation, case, accents, appositive removal
- Fuzzier Matching
  - Dice score (character uni/bi/tri-grams), Hamming, Recursive LCS/iSubstring, Subsequences
  - Word removal (e.g., Inc., US) and abbrev.

Entity Linking: Document Features

- BoW Comparisons
  - TF/IDF & Dice scores for news article and KB text
  - Examined entire articles and passages around query mentions
- Named-Entities
  - Ran BBN's SERIF analyzer on articles
  - Checked for coverage of (1) query co-references and (2) all names/nominals in KB text
  - Noted type, subtype of query entity (e.g., ORG/Media)
- KB Facts
  - Looked to see if candidate node's attributes are present in article text (e.g., spouse, employer, nationality)
- Wiktology
  - UMBC system predicts relevant Wikipedia pages (or KB nodes) for text

Question Answering

Question Answering: Ambiguity
More complication: Opinion Question Answering

Q: What is the international reaction to the re-election of Robert Mugabe as President of Zimbabwe?

A: African observers generally approved of his victory while Western Governments strongly denounced it.

Stoyanov, Cardie, Wiebe 2005
Somasundaran, Wilson, Wiebe, Stoyanov 2007

Subjectivity and Sentiment Analysis

- The linguistic expression of somebody’s opinions, sentiments, emotions, evaluations, beliefs, speculations (private states)

<table>
<thead>
<tr>
<th>Sentiment Analysis</th>
<th>Subjectivity analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Subjective</td>
</tr>
<tr>
<td>Negative</td>
<td>Objective</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
</tr>
</tbody>
</table>

Subjectivity analysis classifies content in objective or subjective

Stoyanov, Cardie, Wiebe 2005
Somasundaran, Wilson, Wiebe, Stoyanov 2007

Sentiment classification
- Document level
- Sentence level
- Product feature level
  - “For a heavy pot, the handle is not well designed.”
- Find opinion holders and their opinions

Subjectivity & Sentiment: More Applications

Product review mining: Best Android phone in the market?
Sentiment tracking

Tracking sentiments toward topics over time:
Is anger ratcheting up or cooling down?

Source: Research.ly

Sentiment Analysis Resources:
Lexicons

Rao & Ravichandran, 2009

Sentiment Analysis Resources:
Corpora

- Pang and Lee, Amazon review corpus
- Blitzer, multi-domain review corpus
Dependency Parsing: Constraints

- Commonly imposed constraints:
  - Single-head (at most one head per node)
  - Connectedness (no dangling nodes)
  - Acyclicity (no cycles in the graph)
  - Projectivity:
    - An arc \( i \rightarrow j \) is projective iff, for every \( k \) occurring between \( i \) and \( j \) in the input string, \( i \rightarrow j \).
    - A graph is projective iff every arc in \( A \) is projective.

Dependency Parsing: Approaches

- Link grammar (Sleator and Temperley)
- Bilexical grammar (Eisner):
  - Lexicalized parsing in \( O(n^3) \) time
- Maximum Spanning Tree (McDonald)
- CONLL 2006/2007

Semantic Role Labeling

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  \[
  \begin{align*}
  \text{agent} & \quad \text{patient} \quad \text{source} \quad \text{destination} \quad \text{instrument} \\
  \text{John} & \quad \text{drove} \quad \text{Mary} \quad \text{from} \quad \text{Austin} \quad \text{to} \quad \text{Dallas} \quad \text{in} \quad \text{his} \quad \text{Toyota Prius}. \\
  \text{The} & \quad \text{hammer} \quad \text{broke} \quad \text{the} \quad \text{window}.
  \end{align*}
  \]

- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”

SRL Datasets

- FrameNet:
  - Developed at UCB
  - Based on notion of Frames
- PropBank:
  - Developed at UPenn
  - Based on elaborating the Treebank
- Salsa:
  - Developed at Universität des Saarlandes
  - German version of FrameNet

SRL as Sequence Labeling

- SRL can be treated as an sequence labeling problem.
- For each verb, try to extract a value for each of the possible semantic roles for that verb.
- Employ any of the standard sequence labeling methods
  - Token classification
  - HMMs
  - CRFs
SRL with Parse Trees

- Parse trees help identify semantic roles through exploiting syntactic clues like "the agent is usually the subject of the verb".
- Parse tree is needed to identify the true subject.

```
                   s
                  /   |
     Det NPsg NPs s
      /        /     |
   NPpl Prep NPP NPp

"The man by the store near the dog ate an apple."
"The man" is the agent of "ate" not "the dog."
```

Selectional Restrictions

- **Selectional restrictions** are constraints that certain verbs place on the filler of certain semantic roles.
  - Agents should be animate
  - Beneficiaries should be animate
  - Instruments should be tools
  - Patients of "eat" should be edible
  - Sources and Destinations of "go" should be places.
  - Sources and Destinations of "give" should be animate.

- Taxonomic abstraction hierarchies or ontologies (e.g. hypernym links in WordNet) can be used to determine if such constraints are met.
  - "John" is a "Human" which is a "Mammal" which is a "Vertebrate" which is an "Animate"

Word Senses

- **Beware of the burning coal underneath the ash.**

  **Ash**
  - Sense 1
  - Sense 2
  - Sense 3

  **Coal**
  - Sense 1
  - Sense 2
  - Sense 3

Self-training via Yarowsky’s Algorithm

Recognizing Textual Entailment

- **Expected answer form**
  - Who bought Overture? >> X bought Overture

- **Hypothesized answer**
  - Overture’s acquisition by Yahoo
  - Yahoo bought Overture

- **Similiar for IE:** X acquire Y
- **Similiar for “semantic” IR**
- **Summarization (multi-document)**
- **MT evaluation**

(Statistical) Machine Translation

- **Language Model**
  - source P(e)
  - decoder

- **Translation Model**
  - channel P(f | e)
  - observed f
Where will we get $P(F|E)$?

Books in English  Same books, in French

We call collections stored in two languages **parallel corpora** or **parallel texts**

Want to update your system? Just add more text!

*Where Next?*

---

**Machine Translation**

- **Systems**
  - Early rule based systems
  - Word based models (IBM models)
  - Phrase based models (log-linear!)
  - Tree based models (syntax driven)
  - Adding semantics (WSD, SRL)
  - Ensemble models

- **Evaluation**
  - Metrics (BLEU, BLACK, ROUGE ...)
  - Corpora (statmt.org)

---

**Allied Areas and Tasks**

- Information Retrieval
  - TREC (Large scale experiments)
  - CLEF (Cross Lingual Evaluation Forum)
  - NTIR
  - FIRE (South Asian Languages)

---

**Allied Areas and Tasks**

- (Computational) Musicology
  - MIREX

---

**Where Next?**