Manual Annotation for NLP

Razvan C. Bunescu
Department of Computer Science @ CCI

razvan.bunescu@uncc.edu
Supervised Learning

• State-of-the-art performance in NLP is obtained using ML models.
• ML models typically require labeled training examples:
  – For sentiment analysis, need documents annotated with sentiment labels.
  – Sentiment analysis was modeled as a classification task.
  – Many other tasks in NLP can be modeled as classification.

• Information Extraction:
  – Named Entity Recognition (NER) as classification.
  – Relation Extraction (RE) as classification.
  – Coreference resolution (Coref) as classification.
    • Mention Detection as classification.
  – Named Entity Disambiguation (NED) as ranking.
Information Extraction: A Pipeline Approach

1. Raw text
2. Tokenization
3. NE Recognition
4. NE Disambiguation
5. Relation Extraction
6. Mention Detection
7. Coreference Resolution

Wikipedia
Information Extraction: Syntax is Useful

Raw text → Tokenization → POS Tagging → Syntactic Parsing

Wikipedia

NE Recognition → NE Disambiguation
Relation Extraction
Mention Detection → Coreference Resolution
Protein name recognition:

Interferon beta was found to upregulate IL-15 in vitro

Labels for protein name recognition:
- B-Prot indicates a token that starts a protein name.
- I-Prot indicates a token that is inside a protein name.
- O indicates any other token.

B-Prot I-Prot O O O O B-Prot O O

Interferon beta was found to upregulate IL-15 in vitro
Named Entity Recognition (NER)

- Protein name recognition:
  
  B-Prot  I-Prot  O  O  O  O  B-Prot  O  O  
  Interferon beta was found to upregulate IL-15 in vitro

- What kind of Boolean features would be useful?
  - Word: Is the current token $w_0$ the name of a Greek letter?
  - Suffix: Does $w_0$ end with a number? Does $w_0$ end with ‘gen’? Does $w_0$ start with an acronym? …
  - Lexicon: Is $w_0$ an entry in a dictionary of known protein names?

  **Word features:**
  - Word identity.
  - Prefix/suffix.
  - Capitalization.
  - Word ‘shape’.
  - Word clusters.

  **Context features:**
  - Words before / after.

  **Lexicon:**
  - An entry.
  - First token in an entry.
Manual Annotation for NE Recognition

- **Internal** annotation:
  - Insert information in the tokenized text itself to indicate the labels.
  - There are many ways in this can be done:
    1. Use XML tags to indicate names.
      
      \(<\textit{prot}>\text{Interferon beta}</\textit{prot}>\) was found to upregulate \(<\textit{prot}>\text{IL-15}</\textit{prot}>\) in vitro
    
    2. Use special characters to indicate labels, e.g. / or _:
      
      Interferon\textit{B} beta\textit{I} was\textit{O} found\textit{O} to\textit{O} upregulate\textit{O} IL-15\textit{B} in\textit{O} vitro\textit{O}
Manual Annotation for NE Recognition

- **External annotation:**
  - Assign a unique position to each token (token-level) or character (character-level) in the input text.
  - Indicate the named entity spans as pairs `<begin-position, end-position>` for each name, in a separate text file.

```
B-Prot  I-Prot  O   O   O   O   B-Prot  O   O
```

**input.txt:**
```
Interferon beta was found to upregulate IL-15 in vitro
```

**token positions**
```
0  1  2  3  4  5  6  7  8
```

**annotation.txt:**
```
0  2
6  7
```
Relation Extraction (RE) as Classification

- If sentence contains $n = 3$ names $\text{name}_1$, $\text{name}_2$, $\text{name}_3$ in this order, and only $\text{name}_1$ and $\text{name}_2$ are annotated to be in a relationship:

  \[ \text{relation } R \]

  \[ \ldots \text{name}_1 \ldots \text{name}_2 \ldots \text{name}_3 \ldots \]

  - Create $(n \choose 2) = 3$ relation examples from each pair, one positive and two negative:
    - Label 1: “\ldots <p>\text{name}_1</p> \ldots \text{name}_2 \ldots \text{name}_3 \ldots”
    - Label 0: “\ldots <p>\text{name}_1</p> \ldots \text{name}_2 \ldots <p>\text{name}_3</p> \ldots”
    - Label 0: “\ldots \text{name}_1 \ldots \text{name}_2 \ldots <p>\text{name}_3</p> \ldots”
  
  - Create features that depend on the position of the two names in the pair:
    - Does a certain word, e.g. “upregulates”, appear between the two names?
    - If syntax available, can create features from path of syntactic dependence between the 2 names.
**Manual Annotation for Relation Extraction (RE)**

- **Internal** annotation as attributes on entity mention annotations:
  - Use XML tags to indicate names, insert attribute to uniquely identify each name.
  - Use relation name as attribute for the last name in the relationship.

  `<prot id="1">Interferon beta</prot> was found to upregulate <prot id="2" ppi="1">IL-15</prot> in vitro.
  - ppi =1 indicates this protein name is in a PPI relation with the protein with id 1.

- **External** annotation by recording pairs of name id’s in a separate text file:
  - input.txt:

```plaintext
<prot id="1">Interferon beta</prot> was found to upregulate <prot id="2">IL-15</prot> in vitro.
```

  - annotations.txt
    ```plaintext
    1 2
    ```
For sentiment analysis, all examples were included in one file:
- Label followed by the text of the movie review.

For general document classification, use external annotation:
- Leave documents in a folder called “data”.
- Have a separate file “corpus.txt” that lists for each document its label.
- Better for large documents, easier to map annotated label to document manually.

```
data/
document_1

corpus.txt
  1 document_1
  0 document_2
  1 document_3
  ...
  0 document_N
```
This paper reviews 78 annotation tools:

You are free to use any tool you like.

Some of them are Web-bases, such as Brat:
- It can be downloaded from [https://webpages.charlotte.edu/rbunescu/courses/itcs4111/brat.zip](https://webpages.charlotte.edu/rbunescu/courses/itcs4111/brat.zip).
- Installation instructions are [here](https):
  - Install standalone server: `./install.sh --u`
  - Start standalone server: `python standalone.py` (must be python 2!)
    - Point the browser to the `address:port` shown as output:
      » Serving brat at http://127.0.0.1:8001
Manual Text Annotation with Brat: NE REcognition

- More details at http://brat.nlplab.org/introduction.html
Manual Text Annotation with Brat: Relation Extraction

Manual Text Annotation with Brat: NE Disambiguation

- More details at http://brat.nlplab.org/introduction.html
Manual Text Annotation with Brat

- Brat records **external** annotations in an `.ann` file.
  - See example files in brat/data/tutorials/mytask/
    - `annotation.conf` is edited to include our NE types (school, protein) and RE types (Attend, PPI).
    - `markov.txt` contains the tokenized input.
    - `markov.ann` records the external annotations.

```entities
# Definition of entities.
# Format is a simple list with one type per line.
Person
Organization
GPE
Money
School
Protein
```

```events
# Definition of events.
# Format in brief: one event per line, with first space-separated
# field giving the event type and the rest of the line the
# comma-separated arguments in ROLE:TYPE format. Arguments may be
# specified as either optional (by appending "?" to role) or repeated
# by appending either "+" for "0 or more" or "+1" for "1 or more").

# this is a macro definition, used for brevity
<POG>=Person|Organization|GPE
# the "!" before a type specifies that it cannot be used for annotation
# (hierarchy structure only.)

Life
Be-born Person-Arg:Person, Place-Arg?:GPE
Marry Person-Arg{2}:Person, Place-Arg?:GPE
Divorce Person-Arg{2}:Person, Place-Arg?:GPE
Die Person-Arg:Person, Agent-Arg?:<POG>, Place-Arg?:GPE
```

```relations
# Definition of (binary) relations.
# Format in brief: one relation per line, with first space-separated
# field giving the relation type and the rest of the line the
# comma-separated arguments in ROLE:TYPE format. The roles are
# typically "Arg1" and "Arg2".

Located Arg1:Person, Arg2:GPE
Geographical_part Arg1:GPE, Arg2:GPE
Family Arg1:Person, Arg2:Person
Employment Arg1:Person, Arg2:GPE
Ownership Arg1:Person, Arg2:Organization
Origin Arg1:Organization, Arg2:GPE
Alias Arg1:Person, Arg2:Person, <REL-TYPE>:symmetric-transitive
Attend_school Arg1:Person, Arg2:School
PPI Arg1:Protein, Arg2:Protein
```
Manual Text Annotation with Brat

• Annotation is done in the browser using the mouse to select and drag text spans.

NER / Mention Detection:
• Protein
• Person
• GPE
• School

RE / Events:
• PPI
• Be born
• Attend school

Coreference relations:
• Alias
Manual Text Annotation with Brat

- Brat records **external** annotations in an `.ann` file, e.g. `markov.ann`:
  - Spans are recorded as character-level positions in the input file, e.g. `markov.txt`.

<table>
<thead>
<tr>
<th>T1</th>
<th>Protein 0 15</th>
<th>Interferon beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>Protein 44 57</td>
<td>ISG95 protein</td>
</tr>
<tr>
<td>R1</td>
<td>PPI Arg1:T1 Arg2:T2</td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>Be-born 81 85</td>
<td>born</td>
</tr>
<tr>
<td>E1</td>
<td>Be-born:T3 Person–Arg:T4 Place–Arg:T5</td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>Person 63 76</td>
<td>Andrey Markov</td>
</tr>
<tr>
<td>T5</td>
<td>GPE 105 111</td>
<td>Russia</td>
</tr>
<tr>
<td>T6</td>
<td>School 130 159</td>
<td>St. Petersburg Grammar School</td>
</tr>
<tr>
<td>T7</td>
<td>Person 114 116</td>
<td>He</td>
</tr>
<tr>
<td>R2</td>
<td>Attend_school Arg1:T7 Arg2:T6</td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>Alias T7 T4</td>
<td></td>
</tr>
</tbody>
</table>
Homework Assignment

1. Propose a custom classification task and create a corpus of documents that are annotated for this task.
   – You can use one of the annotation schemes or tools discussed above.
   – There should be at least 300 examples in total.

2. Propose discriminative features to be included in the feature vector representation for the examples in this task and evaluate their utility by training and testing Logistic Regression models.
   – Design at least 2 features that are task specific and demonstrate understanding of the task.

3. Run ablation studies.

4. Plot ROC or PR curves.