Information Extraction
Lecture 4 – Named Entity Recognition II

CIS, LMU München
Winter Semester 2023-2024

Prof. Dr. Alexander Fraser, CIS
Seminar topics

• Today is the last day for seminar topics!
Reading

• Please read Sarawagi Chapter 3 for next time
  • Sarawagi talks about classifier based IE in Chapter 3
  • Unfortunately, the discussion is very technical. I would recommend reading it, but not worrying too much about the math (yet), just get the basic idea
  • Please plan to reread Chapter 3 again after we discuss machine learning
Outline

• Topics from last time
  • Evaluation metrics in more detail
  • Quick review of Rule-Based NER
• Evaluations and gold standards in IE
  • Issues in Evaluation of IE
  • Human Annotation for NER
• IE end-to-end
• Introduction: named entity detection as a classification problem
Recall

Measure of how much relevant information the system has extracted (coverage of system).

Exact definition:

Recall = 1 if no possible correct answers

else:

# of correct answers given by system

total # of possible correct answers in text
Precision

Measure of how much of the information the system returned is correct (accuracy).

Exact definition:

\[
\text{Precision} = \begin{cases} 
1 & \text{if no answers given by system} \\
\frac{\# \text{ of correct answers given by system}}{\# \text{ of answers given by system}} & \text{else:}
\end{cases}
\]
Evaluation

Every system, algorithm or theory should be evaluated, i.e. its output should be compared to the gold standard (i.e. the ideal output). Suppose we try to find scientists...

Algorithm output:
$$O = \{ \text{Einstein, Bohr, Planck, Clinton, Obama} \}$$

Gold standard:
$$G = \{ \text{Einstein, Bohr, Planck, Heisenberg} \}$$

Precision:
What proportion of the output is correct?
$$\frac{|O \cap G|}{|O|}$$

Recall:
What proportion of the gold standard did we get?
$$\frac{|O \cap G|}{|G|}$$
Evaluation

• Why Evaluate?
• What to Evaluate?
• How to Evaluate?
Why Evaluate?

- Determine if the system is useful
- Make comparative assessments with other methods/systems
  - Who’s the best?
- Test and improve systems
- Others: Marketing, …
What to Evaluate?

- In Information Extraction, we try to match a pre-annotated gold standard
- But the evaluation methodology is mostly taken from Information Retrieval
  - So let's consider *relevant documents* to a search engine *query* for now
  - We will return to IE evaluation later
Relevant vs. Retrieved Documents

All docs available

Set approach
Contingency table of relevant and retrieved documents

- Precision: \( P = \frac{\text{Ret}_\text{Rel}}{\text{Ret}} \)
- Recall: \( R = \frac{\text{Ret}_\text{Rel}}{\text{Relevant}} \)

Total # of documents available \( N = \text{Ret}_\text{Rel} + \text{NotRet}_\text{Rel} + \text{Ret}_\text{NotRel} + \text{NotRet}_\text{NotRel} \)

Slide from Giles
Contingency table of classification of documents

Test result

<table>
<thead>
<tr>
<th>Actual Condition</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>Negative</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

Total # of cases $N = tp + fp + fn + tn$

fp type 1 error

fn type 2 error

present $= tp + fn$

positives $= tp + fp$

negatives $= fn + tn$

Slide modified from Giles
<table>
<thead>
<tr>
<th>Test result</th>
<th>Actual condition</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Condition Present + Positive result = True Positive</td>
<td>Condition absent + Positive result = False Positive (Type I error)</td>
</tr>
<tr>
<td>Negative</td>
<td>Condition present + Negative result = False (invalid) Negative (Type II error)</td>
<td>Condition absent + Negative result = True (accurate) Negative</td>
</tr>
</tbody>
</table>

Example, using infectious disease test results:

<table>
<thead>
<tr>
<th>Test result</th>
<th>Actual condition</th>
<th>Not infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test shows &quot;infected&quot;</td>
<td>True Positive</td>
<td>False Positive (Type I error)</td>
</tr>
<tr>
<td>Test shows &quot;not infected&quot;</td>
<td>False Negative (Type II error)</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

Example, testing for guilty/not-guilty:

<table>
<thead>
<tr>
<th>Test result</th>
<th>Actual condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verdict of &quot;guilty&quot;</td>
<td>True Positive</td>
</tr>
<tr>
<td>Verdict of &quot;not guilty&quot;</td>
<td>False Negative (Type II error)</td>
</tr>
</tbody>
</table>

Example, testing for innocent/not innocent – sense is reversed from previous example:

<table>
<thead>
<tr>
<th>Test result</th>
<th>Actual condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judged &quot;innocent&quot;</td>
<td>True Positive</td>
</tr>
<tr>
<td>Judged &quot;not innocent&quot;</td>
<td>False Negative (Type II error)</td>
</tr>
</tbody>
</table>
Retrieval example

- Documents available: D1, D2, D3, D4, D5, D6, D7, D8, D9, D10
- Relevant: D1, D4, D5, D8, D10
- Query to search engine retrieves: D2, D4, D5, D6, D8, D9

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>not retrieved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Retrieval example

- Documents available: D1, D2, D3, D4, D5, D6, D7, D8, D9, D10
- Relevant: D1, D4, D5, D8, D10
- Query to search engine retrieves: D2, D4, D5, D6, D8, D9

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>D4, D5, D8</td>
<td>D2, D6, D9</td>
</tr>
<tr>
<td>not retrieved</td>
<td>D1, D10</td>
<td>D3, D7</td>
</tr>
</tbody>
</table>
Contingency table of relevant and retrieved documents

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>Ret\text{Rel} = 3</td>
<td>Ret\text{NotRel} = 3</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>NotRet\text{Rel} = 2</td>
<td>NotRet\text{NotRel} = 2</td>
</tr>
</tbody>
</table>

- Relevant documents = Ret\text{Rel} + NotRet\text{Rel} = 3 + 2 = 5
- Not Relevant documents = Ret\text{NotRel} + NotRet\text{NotRel} = 2 + 2 = 4
- Total # of docs N = Ret\text{Rel} + NotRet\text{Rel} + Ret\text{NotRel} + NotRet\text{NotRel} = 10

- Precision: \( P = \frac{\text{Ret\text{Rel}}}{\text{Retrieved}} = \frac{3}{6} = .5 \)
- Recall: \( R = \frac{\text{Ret\text{Rel}}}{\text{Relevant}} = \frac{3}{5} = .6 \)

Slide modified from Giles
What do we want

• Find everything relevant – high recall
• Only retrieve what is relevant – high precision
Relevant vs. Retrieved

All docs

Relevant

Retrieved
Precision vs. Recall

Precision = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}

Recall = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}

All docs

Retrieved

Relevant
Retrieved vs. Relevant Documents

Very high precision, very low recall
Retrieved vs. Relevant Documents

High recall, but low precision
Retrieved vs. Relevant Documents

Very low precision, very low recall (0 for both)

Slide from Giles
Retrieved vs. Relevant Documents

High precision, high recall (at last!)
Why Precision and Recall?

Get as much of what we want while at the same time getting as little junk as possible.

Recall is the percentage of relevant documents returned compared to everything that is available!

Precision is the percentage of relevant documents compared to what is returned!

The desired trade-off between precision and recall is specific to the scenario we are in.
Relation to Contingency Table

- Accuracy: \(\frac{a+d}{a+b+c+d}\)
- Precision: \(\frac{a}{a+b}\)
- Recall: \(\frac{a}{a+c}\)
- Why don’t we use Accuracy for IR? (Assuming a large collection)
  - Most docs aren’t relevant
  - Most docs aren’t retrieved
  - Inflates the accuracy value

<table>
<thead>
<tr>
<th></th>
<th>Doc is Relevant</th>
<th>Doc is NOT relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc is retrieved</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Doc is NOT retrieved</td>
<td>(c)</td>
<td>(d)</td>
</tr>
</tbody>
</table>
CMU Seminars task

• Given an email about a seminar
• Annotate
  – Speaker
  – Start time
  – End time
  – Location
This Monday, 4/26, Prof. Makoto Nagao will give a seminar in the CMT red conference room on recent MT research results.
Creating Rules

• Suppose we observe "the seminar at <stime>4 pm</stime> will [...]" in a training document
• The processed representation will have access to the words and to additional knowledge
• We can create a very specific rule for <stime>
  – And then generalize this by dropping constraints (as discussed previously)
Example

the seminar at <time> 4 pm will

<table>
<thead>
<tr>
<th>Condition</th>
<th>Additional Knowledge</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Lemma</td>
<td>LexCat</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td>Art</td>
</tr>
<tr>
<td>seminar</td>
<td>Seminar</td>
<td>Noun</td>
</tr>
<tr>
<td>at</td>
<td>at</td>
<td>Prep</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Digit</td>
</tr>
<tr>
<td>pm</td>
<td>pm</td>
<td>Other</td>
</tr>
<tr>
<td>will</td>
<td>will</td>
<td>Verb</td>
</tr>
</tbody>
</table>
Example

the seminar at `<time>` 4 pm will

<table>
<thead>
<tr>
<th>Condition</th>
<th>Additional Knowledge</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Lemma</td>
<td>LexCat</td>
</tr>
<tr>
<td>at</td>
<td>at</td>
<td>Prep</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Digit</td>
</tr>
<tr>
<td>pm</td>
<td>pm</td>
<td>Other</td>
</tr>
</tbody>
</table>

`stime`
Example

the seminar at \textless time\textgreater 4 pm will

<table>
<thead>
<tr>
<th>Condition</th>
<th>Additional Knowledge</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Lemma</td>
<td>LexCat</td>
</tr>
<tr>
<td>at</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digit</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For each rule, we look for:

- Support (training examples that match this pattern)
- Conflicts (training examples that match this pattern with no annotation, or a different annotation)

Suppose we see:

"tomorrow at <stime>9 am</stime>"

- The rule in our example applies!
- If there are no conflicts, we have a more general rule

Overall: we try to take the most general rules which don't have conflicts
Returning to Evaluation

• This time, evaluation specifically for IE
Importance of Evaluation in IE

- IE was born from a series of competitive evaluations organised by DARPA in the US
  - MUC Conferences, 1989-1998
    - IE as a departure from IR but using the same types of measures of accuracy
    - The idea was to understand what worked and what not in text analysis
      - Finding a way to compare IE systems and approaches in a controlled way

- Evaluation is in IE’s DNA
  - Publishing IE papers without evaluation is not considered acceptable
Organising Evaluation

• You will need:
  • An annotated training corpus
    • That you will use to develop rules or to train a machine learning algorithm
  • A result scorer
    • A tool that automatically computes accuracy of the system against an annotated corpus
    • E.g. The MUC Scorer
  • An annotated test corpus
    • To be used blindly to test results
      • Please note that run on test corpus should be a one off test
      • Test corpus is not be used to fine tuning accuracy in any way
      • E.g. By looking at the results and changing your rules or by tuning the learning parameters
The Rationale Behind

- **Precision**: how correct is the average answer provided by the system
- **Recall**: how many (correct) pieces of information are retrieved by the system
- **F-measure**: allows comparative evaluations
Evaluation Measures

\[
\text{Recall} = \frac{\text{CORRECT} + (\text{PARTIAL} \times 0.5)}{\text{POSSIBLE}}
\]

\[
\text{Precision} = \frac{\text{CORRECT} + (\text{PARTIAL} \times 0.5)}{\text{ACTUAL}}
\]

\[
F(\beta) = \frac{(\beta^2 + 1) \times \text{PREC} \times \text{REC}}{\beta^2 \times \text{PREC} + \text{REC}}
\]

F-Measure is to be used to compare systems in all evaluations all the three measures must be published.
False Negative in CMU Seminars

• Gold standard test set:

  Starting from <stime>11 am</stime>

• System marks nothing:

  Starting from 11 am

• False negative (which measure does this hurt?)
False Positive in CMU Seminars

• Gold standard test set:
  
  ... Followed by lunch at 11:30 am, and meetings

• System marks:

  ... at $\langle$stime$\rangle$11:30 am$\langle$/stime$\rangle$

• False positive (which measure does this hurt?)
Mislabeled in CMU Seminars

• Gold standard test set:
  at a different time - <stime>6 pm</stime>

• System marks:
  ... - <etime>6 pm</etime>

• What sort of error do we have here?
• Which measures are affected?
• Note that this is different from Information Retrieval!
Partial Matches in CMU Seminars

- Gold standard test set:
  ... at \langle\text{stime}\rangle5\ \text{pm}\langle/\text{stime}\rangle

- System marks:
  ... at \langle\text{stime}\rangle5\langle/\text{stime}\rangle\ \text{pm}

- Then I get a partial match (worth 0.5)
- Also different from Information Retrieval
Issues in Evaluation
Issues Affecting Evaluation

- The Algorithm
- The feature set used
- The leniency in assessing results
  - the availability of standard annotated corpora do not guarantee that the experiments performed with different approaches and algorithms proposed in the literature can be reliably compared
    - Data problems
    - Problems of experimental design
    - Problems of presentation

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliani, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
Leniency in Evaluation

- **Data Problems**
  - Errors in data, branching corpora, templates Vs markup

- **Experimental design**
  - Training/Test Set selection
    - e.g. 50/50 Vs 80/20
  - Tokenization
  - How to count matches (see below)

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
Issues in Evaluation

• Fragment evaluation:
  • How leniently should inexact identification of filler boundaries be assessed?

• Counting multiple matches:
  • When a learner predicts multiple fillers for an entity, how should they be counted?

• Filler variation:
  • When text fragments having distinct surface forms refer to the same underlying entity, how should they be counted?

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliani, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
• Evaluation is a critical issue where there is still much work to be done
• But before we can evaluate, we need a **gold standard**
• Training IE systems
  – Critical component for "learning" statistical classifiers
  – The more data, the better the classifier
• Can also be used for developing a handcrafted NER system
  – Constant rescoring and coverage checks are very helpful
• Necessary in both cases for **evaluation**
Annotating Documents to IE Train Systems

Can we really ask people to annotate documents?

Most slides are from Ziqi Zhang, University of Sheffield
Do People Like Annotating?

• No, they hate it
  • They will try not to do it or do it quickly
• It is time and energy consuming
  • It is not their job
    • Unless they are professional annotators
  • They are not rewarded for it
• It is tiring
• It is error prone
• But most of all: is it possible to annotate documents with sufficient accuracy to train an IE system?
A project funded by AHRC/EPSRC/JISC in the UK. In collaboration with the University of York (Archaeology Department)

Goal:

- Building an e-archaeology application to allow archaeologists to discover, share, and analyse datasets and legacy publications

Role of IE: To identify in several collections of documents:

- Pacenames: around 2,000 in corpus
- Subjects: around 10,000
  - Roman pottery, spearhead, animal remains, church, courtyard, plates, vessel
- Temporals: around 4,000
  - Roman, Saxon, AD1078, 300BC, 43 - 801AD, circa 1771, Victorian era, Bronze Age

http://nlp.shef.ac.uk/wig/research/ArchaeoTools.html

Wednesday, 26 August 2009
IE in Aracheotools

• Based on SVN
  • The TRex tool http://t-rex.sourceforge.net/

• Training based on corpora annotated by 5 expert archaeologists
  • training documents 42, length: up to several hundreds of pages
  • total documents to tag by machine learning: 967
  • total documents to tag by rules: 3991

• Annotation process was geared at high quality
  • Annotation instructions were clarified through several iterations
    • Our archaeologists colleagues, they clearly explained the task to annotators, went through examples with them
    • The IE experts went through several confusing examples with archaeologists to clarify their doubts
  • One senior researcher was appointed to make final decision in case of doubts from any annotators
  • Annotators were very motivated and the task was part of their job!!!
IE challenges – annotation quality


<table>
<thead>
<tr>
<th></th>
<th>Annotator A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Annotator B</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
</tr>
</tbody>
</table>

- Treating A’s annotations as gold standard, and B’s as reference
- Precision of B = a/(a+b), Recall of B = a/(a+c)
- F-measure of B = 2a/(2a+b+c)
- Equivalent to the standard P, R, F metrics used for evaluating IE systems
Annotation quality (ctd.)

- IAA F-measure – Inter-Annotator-Agreement F-measure

✓ Figures obtained from a shared corpus annotated by three different annotators

<table>
<thead>
<tr>
<th></th>
<th>Place name</th>
<th>Subject</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest IAA between any two annotators</td>
<td>66.2</td>
<td>49</td>
<td>67.2</td>
</tr>
<tr>
<td>Highest IAA between any two annotators</td>
<td>80</td>
<td>63</td>
<td>83.3</td>
</tr>
</tbody>
</table>
Annotator Variability

- Differences in annotation are a significant problem
  - Only some people are good at annotation
  - Practice helps

- Even good annotators can have different understanding of the task
  - For instance, in doubt, annotate? Or not?
  - (~ precision/recall tradeoffs)

- Effect of using gold standard corpora that are not well annotated
  - Evaluations can return inaccurate results
  - Systems trained on inconsistent data can develop problems which are worse than if the training examples are eliminated

- Crowd-sourcing, which we will talk about later, has all of these same problems even more strongly!
Annotation Quality - Conclusions

• In general archaeology is a difficult domain, with many uncertainty and ambiguity even for humans

• Inconsistency between annotators generated noise that influences learning system

• Very careful evaluation of the quality of annotation must always be implemented
  • Aka possibility/ability for the annotators to perform good quality annotation

• Never ever suppose that humans are 100% correct
  • For complex tasks they may perform at 80% accuracy!!!!
    • Always ask users to annotate (at least partially) overlapping sets of documents
      • So to be able to check their agreement
CMU Seminars task

- Given an email about a seminar…
- Annotate mentions of:
  - Speaker
  - Start time
  - End time
  - Location
This Monday, 4/26, Prof. Makoto Nagao will give a seminar in the CMT red conference room on recent MT research results.
# IE Template

<table>
<thead>
<tr>
<th>Slot Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker</td>
<td>Prof. Makoto Nagao</td>
</tr>
<tr>
<td>Start time</td>
<td>1993-04-26 10:00</td>
</tr>
<tr>
<td>End time</td>
<td>1993-04-26 11:00</td>
</tr>
<tr>
<td>Location</td>
<td>CMT red conference room</td>
</tr>
<tr>
<td>Message Identifier (Filename)</td>
<td><a href="mailto:0.24.4.93.20.59.10.jgc+@NL.CS.CMU.EDU">0.24.4.93.20.59.10.jgc+@NL.CS.CMU.EDU</a> (Jaime Carbonell).0</td>
</tr>
</tbody>
</table>

- Template contains *canonical* version of information
- There are several "mentions" of speaker, start time and end-time in the email (see previous slide)
- Only one value for each slot
- Location could probably also be canonicalized
- Important: also keep link back to original text
How many database entries?

• In the CMU seminars task, one message generally results in one database entry
  – Or no database entry if you process an email that is not about a seminar

• In other IE tasks, can get multiple database entries from a single document or web page
  – A page of concert listings -> database entries
  – Entries in timeline -> database entries
Summary

• IR: end-user
  – Start with information need
  – Gets relevant documents, hopefully information need is solved
  – Important difference: Traditional IR vs. Web R

• IE: analyst (you)
  – Start with template design and corpus
  – Get database of filled out templates
    • Followed by subsequent processing (e.g., data mining, or user browsing, etc.)
IE: what we've seen so far

So far we have looked at:

• Source issues (selection, tokenization, etc)
• Extracting regular entities
• Rule-based extraction of named entities
• Learning rules for rule-based extraction of named entities
• We also jumped ahead and looked briefly at end-to-end IE for the CMU Seminars task
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.

- **Source Selection**
- **Tokenization & Normalization**
  - 05/01/67
  - 1967-05-01
- **Named Entity Recognition**
- **Instance Extraction**
  - ...married *Elvis* on 1967-05-01
- **Fact Extraction**
- **Ontological Information Extraction**

- Elvis Presley: singer
- Angela Merkel: politician
Where we are going

- We will stay with the named entity recognition (NER) topic for a while
  - How to formulate this as a machine learning problem (later in these slides)
  - Next time: brief introduction to machine learning
Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, dates, ...) in a text.

Elvis Presley was born in 1935 in East Tupelo, Mississippi.
Extracting Named Entities

Person: Mr. Hubert J. Smith, Adm. McInnes, Grace Chan
Title: Chairman, Vice President of Technology, Secretary of State
Country: USSR, France, Haiti, Haitian Republic
City: New York, Rome, Paris, Birmingham, Seneca Falls
Province: Kansas, Yorkshire, Uttar Pradesh
Business: GTE Corporation, FreeMarkets Inc., Acme
University: Bryn Mawr College, University of Iowa
Organization: Red Cross, Boys and Girls Club
More Named Entities

Currency: 400 yen, $100, DM 450,000
Linear: 10 feet, 100 miles, 15 centimeters
Area: a square foot, 15 acres
Volume: 6 cubic feet, 100 gallons
Weight: 10 pounds, half a ton, 100 kilos
Duration: 10 day, five minutes, 3 years, a millennium
Frequency: daily, biannually, 5 times, 3 times a day
Speed: 6 miles per hour, 15 feet per second, 5 kph
Age: 3 weeks old, 10-year-old, 50 years of age
IE Posed as a Machine Learning Task

- Training data: documents marked up with ground truth
- Extract features around words/information
- Pose as a classification problem
For each position, ask: Is the current window a named entity?

Window size = 1
For each position, ask: Is the current window a named entity?

Window size = 2
Choose certain **features** (properties) of windows that could be important:

- window contains colon, comma, or digits
- window contains week day, or certain other words
- window starts with lowercase letter
- window contains only lowercase letters
- ...
The feature vector represents the presence or absence of features of one content window (and its prefix window and postfix window).
Sliding Windows Corpus

Now, we need a **corpus** (set of documents) in which the entities of interest have been manually labeled.

NLP class: **Wednesday, 7:30am** and Thursday all day, **rm 667**

From this corpus, compute the feature vectors with labels:

<table>
<thead>
<tr>
<th>time</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Nothing   Nothing   Time   Nothing   Location

Slide from Suchanek
Use the labeled feature vectors as training data for Machine Learning.
Sliding Windows Exercise

What features would you use to recognize person names?

Elvis Presley married Ms. Priscilla at the Aladin Hotel.

UpperCase

hasDigit

...
Good Features for Information Extraction

Example word features:
- identity of word
- is in all caps
- ends in “-ski”
- is part of a noun phrase
- is in a list of city names
- is under node X in WordNet or Cyc
- is in bold font
- is in hyperlink anchor
- features of past & future
- last person name was female
- next two words are “and Associates”

contains-question-mark
contains-question-word
ends-with-question-mark
first-alpha-is-capitalized
indented
indented-1-to-4
indented-5-to-10
more-than-one-third-space
only-punctuation
prev-is-blank
prev-begins-with-ordinal
shorter-than-30
<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Capitalized</td>
<td>Character n-gram classifier says string is a person name (80% accurate)</td>
</tr>
<tr>
<td>Is Mixed Caps</td>
<td></td>
</tr>
<tr>
<td>Is All Caps</td>
<td></td>
</tr>
<tr>
<td>Initial Cap</td>
<td></td>
</tr>
<tr>
<td>Contains Digit</td>
<td></td>
</tr>
<tr>
<td>All lowercase</td>
<td></td>
</tr>
<tr>
<td>Is Initial</td>
<td></td>
</tr>
<tr>
<td>Punctuation</td>
<td></td>
</tr>
<tr>
<td>Period</td>
<td></td>
</tr>
<tr>
<td>Comma</td>
<td></td>
</tr>
<tr>
<td>Apostrophe</td>
<td></td>
</tr>
<tr>
<td>Dash</td>
<td></td>
</tr>
<tr>
<td>Preceded by HTML tag</td>
<td></td>
</tr>
<tr>
<td>Word Features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>lists of job titles,</td>
</tr>
<tr>
<td></td>
<td>Lists of prefixes</td>
</tr>
<tr>
<td></td>
<td>Lists of suffixes</td>
</tr>
<tr>
<td></td>
<td>350 informative phrases</td>
</tr>
<tr>
<td>HTML/Formatting Features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>{begin, end, in} x</td>
</tr>
<tr>
<td></td>
<td>{&lt;b&gt;, &lt;i&gt;, &lt;a&gt;, &lt;hN&gt;} x</td>
</tr>
<tr>
<td></td>
<td>{lengths 1, 2, 3, 4, or longer}</td>
</tr>
<tr>
<td></td>
<td>{begin, end} of line</td>
</tr>
</tbody>
</table>
• Slide sources
  – A number of slides were taken from a wide variety of sources (see the attribution at the bottom right of each slide)
  – Some of the slide authors:
    • C. Lee Giles, Penn State
    • Fabio Ciravegna/Ziqi Zhang, Sheffield
    • Dave Kauchak, Pomona College
    • Fabian Suchanek, Telecom ParisTech
  – CMU Seminars task: Dayne Freitag, see also his PhD thesis (Machine Learning for Information Extraction in Informal Domains, CMU, Nov 1998)
Conclusion

• Last two lectures
  – Manually coded rules for NER
  – Learning rules for NER
  – Evaluation
  – Annotation
  – Introduction to classification
    • Sliding windows and features

• Please read Sarawagi Chapter 3!
Questions?
• Thank you for your attention!