Recent pointer to open-source NLP software
Fun: the **Tensorflow Neural Network Playground**

- Configure and run your own neural net for some sample problems
- Choose dataset, input features, hidden layer topology, train/test partition, batch size, # output units, learning rate, activation, regularization (+ rate), problem type
- Click run, epochs increment
- Classification of data
NAMED ENTITY RECOGNITION (NER) AND MACHINE LEARNING

Named Entity Recognition:

1. President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.
Tokenization

• Identify words from a text
• Process punctuation:
  • Identify sentences, abbreviations
• Identify symbols (numbers, addresses, markup codes, special characters)
• Normalize orthography (spelling, caps, hyphenation, etc.)
ABERNETHY, WILLIAM, Wallingford, m. 1673 or 4, Sarah, d. of William Doolittle, had William, and Samuel, and d. 1718, when his two s. admin. on his est. Early this name was writ. Ebenetha, or Abbenatha, acc. Hinman; but in mod. days the descend. use the spell. here giv.

ABINGTON, WILLIAM, Maine, 1642. Coffin.

ABORNE. See Eborne.

ACRERLY, ACCORLEY, or ACRELY, HENRY, New Haven 1640, Stamford 1641 to 53, Greenwich 1656, d. at S. 17 June 1668, wh. is the date of his will. His wid. Ann, was 75 yrs. old in 1662. Haz. II. 246.

ROBERT, Brookhaven, L. I. 1655, adm. freem. of Conn. jurisdict. 1664. See Trumbull, Col. Rec. I. 341,428. SAMUEL, Brookhaven, 1655, perhaps br. of the preced.
Named entity recognition (NER)

- Locate and classify elements (aka rigid designators) into specific categories
  - People
  - Organizations
  - Locations
  - Time expressions
  - Quantities
  - Monetary values
  - Percentages
  - etc.
Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].
Basics of named entity recognition

• Most NER systems take a block of text:
  • Jim bought 300 shares of Acme Corp. in 2006.

• and annotate it with semantic tags
  • <enamex type="person">Jim</enamex> bought
    <numex type="quantity">300</numex> shares of
    <enamex type="organization">Acme Corp.</enamex> in
    <timex type="date">2006</timex>.

• Used by specialized search engines
• Monitor trends in bodies of text (e.g. newswire)
• Enable researchers (biology, genetics, law) to search for answers in large
  body of text
Kofi Atta Annan is a Ghanaian diplomat who served as the seventh Secretary General of the United Nations from January 1, 1997, to January 1, 2007, serving two five-year terms. Annan was the co-recipient of the Nobel Peace Prize in October 2001.

Kofi Annan was born on April 8, 1938, to Victoria and Henry Reginald Annan in Kumasi, Ghana. He is a twin, an occurrence that is regarded as special in Ghanaian culture. Efua Atta, his twin sister, shares the same middle name, which means ‘twin’. As with most Akan names, his first name indicates the day of the week he was born: ‘Kofi’ denotes a boy born on a Friday. The name Annan can indicate that a child was the fourth in the family, but in his family it was simply a name which Annan inherited from his parents.

In 1962, Annan started working as a Budget Officer for the World Health Organization, an agency of the United Nations. From 1974 to 1976, he was the Director of Tourism in Ghana. Annan then returned to work for the United Nations as an Assistant Secretary General in three consecutive positions.
Common annotation tag types

• Enamex
  • People
    • Politician, entertainer, moderator, etc.
  • Locations
    • City, state, country, etc.
  • Organizations
  • Times (dates, time ranges, etc.)

• Medical terms
  • Disease names

• Miscellaneous
  • Product
  • Time
  • Number
NER methods

• Grammar-based
  • Human writes a grammar that the system follows.
  • Very accurate, very time- and labor- intensive

• Machine learning
  • Computer learns how to do the task (w/ annotations)
  • Not as accurate, faster and more efficient

• Hybrid
  • Combines both
  • State of the art
NER (i.e. machine learning) stages

- Annotated corpus (labeling scheme, dev/devtest/test split)
- Feature extractor
- Algorithm selection
  - Mostly supervised learning
- Train, devtest, retrain w/ better features, devtest, etc.
- Final evaluation with test
Feature extraction

• Use different characteristics of a word or phrase
  • Has this word/phrase been seen before
  • Capitalization
  • Position in sentence
  • Word length
  • Context

• 3 different category types
  • Word-level features
  • List lookup features
  • Document & corpus features
Example Feature Extraction

- **Jim bought 300 shares of Acme Corp. in 2006.**
  - Starts with capital – Y
  - First word of sentence – N
  - Contains punctuation – N
  - String length – 4
  - Position – 5 before, 4 after
  - Context
    - Frequent neighbors (prepositions)
Word-level features

- Case
- Punctuation
- Character
- Part of speech
- Semantic roles
- Digit patterns
  - 2 or 4 digit numbers can stand for year
  - 1 or 2 digit can stand for day or month
- Morphology
  - Common prefixes, suffixes
    - Professions often end in –ist (journalist)
    - Nationality and language often end in –ish/-an (Canadian, Bulgarian, Polish)
    - Organization names often include segments such as “tech” and “soft”
- Applying functions to words can encode useful generalization
  - Map uppercase to “A”, lowercase to “a”, punctuation to “-”, and numbers to “0”
    - G.M. > A-A-
    - Machine-223 > Aaaaaaa-000 > Aa-0
- Embeddings
List lookup

- Lists / gazetteers / lexicons / dictionary
- General list
  - General dictionary
- List of entities
  - Organization, first name, continent, etc.
- List of entity cues
  - Typical words in organization name
  - Person title, name prefix, cardinal point

- Most techniques look for an exact word
  - Very inflexible
- Techniques can be employed to make list lookup more flexible
  - Addition or subtraction of morphemes
  - Fuzzy matching with edit distance
  - Normalization with Soundex algorithm
    - First letter of a word plus a three digit code representing its phonetic sound.
Document and corpus features

• Multiple occurrences
  • Anaphora, coreference
• Local syntax
• Meta-information
  • URI, email header, etc
• Corpus frequency
  • Multiword unit permanency
Dictionary / entity cues

- Disambiguation of words in ambiguous positions (sentence-initial capitalization)
- 2677 ambiguous words
  - Common nouns - 1841 out of 1851
  - Named entities - 655 out of 826
- Organization names
  - Todd Associates
  - AutoCom Associates
  - Jennison Associates
IOB Tagging

- [PER Jane Villanueva] of [ORG United], a unit of [ORG United Airlines Holding], said the fare applies to the [LOC Chicago] route.

<table>
<thead>
<tr>
<th>Words</th>
<th>BIO Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>B-PER</td>
</tr>
<tr>
<td>Villanueva</td>
<td>I-PER</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
</tr>
<tr>
<td>United</td>
<td>B-ORG</td>
</tr>
<tr>
<td>Airlines</td>
<td>I-ORG</td>
</tr>
<tr>
<td>Holding</td>
<td>I-ORG</td>
</tr>
<tr>
<td>discussed the</td>
<td>O</td>
</tr>
<tr>
<td>Chicago</td>
<td>O</td>
</tr>
<tr>
<td>route</td>
<td>B-LOC</td>
</tr>
<tr>
<td>.</td>
<td>O</td>
</tr>
</tbody>
</table>
Standard algorithms for NER

- Supervised Machine Learning given a human-labeled training set of text annotated with tags
- Hidden Markov Models
- Conditional Random Fields (CRF)/ Maximum Entropy Markov Models (MEMM)
- Neural sequence models (RNNs or Transformers)
- Large Language Models (like BERT), finetuned
What next?

• NER’s are important, but not the whole story.
• Coreference: associate pronouns with entities
  • Fred took Mary to the store. He bought her flowers.
• Subsequent mentions:
  • Wells Fargo & Co. said Wednesday… … The San Francisco bank…The bank…Wells Fargo…
• Relationships
  • How entities “connect” with one another
  • Birth event: person(name) + date/time + place
  • Graduation event: person(name) + institution + degree + date
Problems with NER

- NER systems for one area may not work well in another area
  - There can be a 20%-40% drop in accuracy when moving from one category to another.
    - Journal articles (1990s)
    - Military dispatch reports
    - Molecular Biology (1998)
    - Bioinformatics, medical research

- Not all words that start with a capital letter are named entities
  - They have too many hamsters.

- No list of named entities can ever be complete
  - Coinage

- Context can be misleading

- Words can be used in novel ways
  - Just google it.
NER shared tasks / competitions

- Most evaluation is done at conferences or contests put on by government, sometimes acting together with contractors or academics.

<table>
<thead>
<tr>
<th>Conference</th>
<th>Acronym</th>
<th>Language(s)</th>
<th>Year(s)</th>
<th>Sponsor</th>
<th>Archive Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACL Special Interest Group in Chinese</td>
<td>SIGHan</td>
<td>Chinese</td>
<td>2006</td>
<td></td>
<td>[8]</td>
</tr>
</tbody>
</table>
CoNLL 2002

- **Languages**
  - Spanish and Dutch

- **Tags**
  - Person (PER)
  - Organization (ORG)
  - Location (LOC)
  - Miscellaneous (MISC)

- **IOB tags**

<table>
<thead>
<tr>
<th></th>
<th>Spanish</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[CNKD2]</td>
<td>81.38%</td>
<td>61.49%</td>
<td>81.39</td>
<td>±1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[F1002]</td>
<td>78.70%</td>
<td>79.40%</td>
<td>79.05</td>
<td>±1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[CV01]</td>
<td>78.14%</td>
<td>78.14%</td>
<td>77.18</td>
<td>±1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[WN02]</td>
<td>75.85%</td>
<td>77.38%</td>
<td>76.61</td>
<td>±1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[BBN02]</td>
<td>77.15%</td>
<td>74.44%</td>
<td>75.76</td>
<td>±1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[TM02]</td>
<td>76.00%</td>
<td>75.55%</td>
<td>75.78</td>
<td>±1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[FM02]</td>
<td>74.12%</td>
<td>73.56%</td>
<td>73.92</td>
<td>±1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Jn02]</td>
<td>74.03%</td>
<td>73.76%</td>
<td>73.69</td>
<td>±1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Ma102]</td>
<td>73.90%</td>
<td>73.39%</td>
<td>73.66</td>
<td>±1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Tsu02]</td>
<td>69.04%</td>
<td>74.12%</td>
<td>71.49</td>
<td>±1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[BU02]</td>
<td>60.53%</td>
<td>67.29%</td>
<td>63.73</td>
<td>±1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[MK02]</td>
<td>56.28%</td>
<td>66.51%</td>
<td>60.97</td>
<td>±1.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>56.27%</td>
<td>56.90%</td>
<td>55.86</td>
<td>±1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Dutch</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[CNKD2]</td>
<td>77.03%</td>
<td>76.29%</td>
<td>77.05</td>
<td>±1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[WN02]</td>
<td>76.95%</td>
<td>73.83%</td>
<td>75.36</td>
<td>±1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[F1002]</td>
<td>78.10%</td>
<td>74.89%</td>
<td>74.99</td>
<td>±1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[BM02]</td>
<td>72.69%</td>
<td>72.45%</td>
<td>72.57</td>
<td>±1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[CV01]</td>
<td>78.03%</td>
<td>71.62%</td>
<td>72.31</td>
<td>±1.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[FM02]</td>
<td>74.01%</td>
<td>68.90%</td>
<td>71.36</td>
<td>±1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Jn02]</td>
<td>72.56%</td>
<td>68.88%</td>
<td>70.57</td>
<td>±1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Jn02]</td>
<td>70.12%</td>
<td>69.26%</td>
<td>69.68</td>
<td>±1.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Ma102]</td>
<td>70.98%</td>
<td>65.50%</td>
<td>68.00</td>
<td>±1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Tsu02]</td>
<td>57.33%</td>
<td>65.02%</td>
<td>60.93</td>
<td>±1.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[MM02]</td>
<td>56.22%</td>
<td>63.24%</td>
<td>59.52</td>
<td>±2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[FP02]</td>
<td>51.54%</td>
<td>47.78%</td>
<td>49.75</td>
<td>±2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>81.29%</td>
<td>75.42%</td>
<td>78.26</td>
<td>±1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Wolff B-PER
- currently O
- a O
- journalist O
- in O
- Argentina B-LOC
- , O
- played O
- with O
- Del B-PER
- Bosque I-PER
- in O
- the O
- final O
- years O
- of O
- the O
- seventies O
- in O
- Real B-ORG
- Madrid I-ORG
- . O
CoNLL 2003

- Languages
  - English and German
- Tags
  - Person (PER)
  - Organization (ORG)
  - Location (LOC)
  - Miscellaneous (MISC)
- IOB tags
- POS tags
- Named Entity Tags

<table>
<thead>
<tr>
<th>U.N.</th>
<th>NNP</th>
<th>I-NP</th>
<th>I-ORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>official</td>
<td>NN</td>
<td>I-NP</td>
<td>O</td>
</tr>
<tr>
<td>Ekeus</td>
<td>NNP</td>
<td>I-NP</td>
<td>I-PER</td>
</tr>
<tr>
<td>heads</td>
<td>VBZ</td>
<td>I-VP</td>
<td>O</td>
</tr>
<tr>
<td>for</td>
<td>IN</td>
<td>I-PP</td>
<td>O</td>
</tr>
<tr>
<td>Baghdad</td>
<td>NNP</td>
<td>I-NP</td>
<td>I-LOC</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Sample results (2003)

- [FIJZ03]
  - used a combination of robust linear classifier, maximum entropy, transformation-based learning, and hidden Markov model
Sample output

- Abner
Asthma and chronic obstructive pulmonary disease (COPD) are chronic airway diseases characterized by airflow obstruction. The beta(2)-adrenoceptor mediates bronchodilatation in response to exogenous and endogenous beta-adrenoceptor agonists. Single nucleotide polymorphisms in the beta(2)-adrenoceptor gene (ADRB2) cause amino acid changes (e.g. Arg16Gly, Gln27Glu) that potentially alter receptor function.
FamilySearch robokey

- 26.5 million obituaries
- NER + relationships
- 1500 years vs. 5 days
NER Software

• Most NLP pipelines (CoreNLP, NLTK, GATE, LingPipe, Spacy, Polyglot, LIMA, etc.)
  • http://cogcomp.cs.illinois.edu/demo/ner/
  • http://en.wikipedia.org/wiki/Named_entity_recognition
  • BYU’s Ontos system
• See also the homework assignment