Information Extraction from Biomedical Text

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Mark Craven
craven@biostat.wisc.edu
Spring 2011

Goals for Lecture

the key concepts to understand are the following
• named-entity recognition (NER) task
• relation extraction task
• sources of evidence for NER
• dictionary based approach to NER
• rule-based approach to NER
• conditional random field representation
• rule-based approach to relation extraction
• event extraction task
The Named Entity Recognition Task

given
passages of text
named-entity classes of interest
  e.g. gene/protein names

recognize
instances of the named entity classes in the text
  e.g. PRP20, SRM1

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the PRP20 gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected. SRM1 mutants, defective in the same gene, influence the signal transduction pathway for the pheromone response . . .

By immunofluorescence microscopy the PRP20 protein was localized in the nucleus. Expression of the RCC1 protein can complement the temperature-sensitive phenotype of PRP20 mutants, demonstrating the functional similarity of the yeast and mammalian proteins

- proteins
- small molecules
- methods
- cellular compartments
The Relation Extraction Task

given
passages of text
relations of interest
e.g. subcellular-localization(Protein, Compartment)
protein-protein-interaction(Protein, Protein)

extract
instances of the relations described in the text
e.g. subcellular-localization(PR20, nucleus)

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the PRP20 gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected. SRM1 mutants, defective in the same gene, influence the signal transduction pathway for the pheromone response . . .

By immunofluorescence microscopy the PRP20 protein was localized in the nucleus. Expression of the RCC1 protein can complement the temperature-sensitive phenotype of PRP20 mutants, demonstrating the functional similarity of the yeast and mammalian proteins

subcellular-localization(PR20, nucleus)
Motivation for Information Extraction

- motivation for named entity recognition
  - better indexing of biomedical articles
  - identifying relevant passages for curation
  - assisting in relation/event extraction

- motivation for relation extraction
  - assisting database curation
  - annotating high-throughput experiments
  - assisting scientific discovery by detecting previously unknown relationships

Aiding Annotation: MGI Example
The Gene Ontology

- a controlled vocabulary of more than 30K concepts describing molecular functions, biological processes, and cellular components

Annotating Genomes: MGI Example

- the current method for this annotation process…
How Do We Get IE Models?

1. encode them by hand
2. learn them from training data

Some Biomedical Named Entity Types

- genes
- proteins
- RNAs
- cell lines/types
- cell components
- diseases/disorders
- drugs
- chromosomal locations
Why Named Entity Recognition is Hard

- these are all gene names
  - CAT1
  - lacZ
  - 3-fucosyl-N-acetyl-lactosamine
  - MAP kinase
  - mitogen activated protein kinase
  - mitogen activated protein kinase kinase
  - mitogen activated protein kinase kinase kinase
  - Hairless
  - onion ring
  - sonic hedgehog
  - And

- in some contexts these names refer to the gene, in other contexts they refer to the protein product, in other contexts its ambiguous

Why Named Entity Recognition is Hard

- they may be referenced conjunctions and disjunctions
  - human B- or T-cell lines
  - human B-cell line         human T-cell line

- there may be variation in orthography
  - NF-kappaB
  - NF KappaB
  - NF-kappa B
  - (NF)-kappaB

- there may be references to gene/protein families
  - OLE1-4
  - OLE1    OLE2    OLE3    OLE4
Identity Uncertainty in NER

• often, there are many names for the same entity

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fut4</td>
<td>fucosyltransferase 4</td>
<td>MG1:5594</td>
</tr>
</tbody>
</table>

| Synonyms                                                                                       |
| 3-fucosyl-N-acetyl-lactosamine, 3-fucosyl-N-acetyl-lactosamine, alpha(1,3) fucosyltransferase, myeloid specific, FAL, FoxT-IV, SSEA-1 |

• synonym lists are often incomplete
• homonymy is also an issue

Sources of Evidence for Biomedical NER

• **orthographic/morphological**: spelling, punctuation, capitalization
  e.g. alphanumeric? contains dashes? capitalized? ends in “ase”
  Src, SH3, p54, SAP, hexokinase

• **lexical**: specific words and word classes
  ___ kinase, ___ receptor, ___ factor

• **syntactic**: how words are composed into grammatical units
  binds to ____, regulated by ____, ____ phosphorylates
Recognizing Protein Names: A Rule-Based Approach
[Fukuda et al., PSB 1998]

1. morphological and lexical analysis is used to identify “core terms” (e.g. Src, SH3, p54, SAP) and “feature terms” (e.g. receptor, protein)

The focal adhesion kinase (FAK) is...

2. lexical and syntactic analysis is used to extend terms into protein names

The focal adhesion kinase (FAK) is...

Recognizing Protein Names: Morphological Analysis in Fukuda Approach

- make list of candidate terms: words that include upper-case letters, digits, and non-alphanumeric characters
- exclude words with length > 9 consisting of lower-case letters and -'s (e.g. full-length)
- exclude words that indicate units (e.g. aa, bp, nM)
- exclude words that are composed mostly of non-alphanumeric characters (e.g. +/-)
Recognizing Protein Names: Lexical/Syntactic Analysis in Fukuda Approach

- merge adjacent terms
  
  Src SH3 domain → Src SH3 domain

- merge non-adjacent terms separated only by nouns, adjectives and numerals
  
  Ras guanine nucleotide exchange factor Sos → Ras guanine nucleotide exchange factor Sos

- extend term to include a succeeding upper-case letter or a Greek-letter word
  
  p85 alpha → p85 alpha
Another Approach:
Dictionaries of Protein Terms
[Bunescu et al., AIM '05]

<table>
<thead>
<tr>
<th>Protein name (OD)</th>
<th>Generalized name (GD)</th>
<th>Canonical form (CD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>interleukin-1 beta</td>
<td>interleukin $\langle y \rangle$</td>
<td>interleukin</td>
</tr>
<tr>
<td>interferon alpha-D</td>
<td>interferon $\langle r \rangle$</td>
<td>interferon</td>
</tr>
<tr>
<td>NF-IL6-beta</td>
<td>NF IL $\langle n \rangle$</td>
<td>NF IL</td>
</tr>
<tr>
<td>TR2</td>
<td>TR $\langle n \rangle$</td>
<td>TR</td>
</tr>
<tr>
<td>NF-kappa B</td>
<td>NF $\langle g \rangle$</td>
<td>NF</td>
</tr>
</tbody>
</table>

- **original dictionary**: extracted 42,172 gene/protein names from HPI and GO databases
- **generalized dictionary**: replaced numbers with $\langle n \rangle$, Roman letters with $\langle r \rangle$, Greek letters with $\langle g \rangle$
- **canonical dictionary**: stripped generic tags from generalized dictionary entries

NER Results from Bunescu et al.

<table>
<thead>
<tr>
<th>IE methods and additional information used</th>
<th>Precision(%)</th>
<th>Recall(%)</th>
<th>F-measure(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original dictionary</td>
<td>56.70</td>
<td>27.24</td>
<td>36.80</td>
</tr>
<tr>
<td>Plus generalized dictionary</td>
<td>62.27</td>
<td>45.85</td>
<td>52.81</td>
</tr>
<tr>
<td>Plus canonical dictionary</td>
<td>41.88</td>
<td>54.42</td>
<td>47.33</td>
</tr>
<tr>
<td>Rmrs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words only</td>
<td>76.11</td>
<td>9.97</td>
<td>17.63</td>
</tr>
<tr>
<td>Part-of-speech</td>
<td>70.84</td>
<td>11.05</td>
<td>19.12</td>
</tr>
<tr>
<td>Dictionary-based tagger</td>
<td>74.49</td>
<td>12.22</td>
<td>21.00</td>
</tr>
<tr>
<td>BWI (300 iterations, 2 lookheads, max. recall)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words only</td>
<td>70.67</td>
<td>11.52</td>
<td>19.81</td>
</tr>
<tr>
<td>Dictionary-based tagger</td>
<td>71.91</td>
<td>24.06</td>
<td>35.94</td>
</tr>
<tr>
<td>k-NN ($k = 1, N = 2$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-of-speech</td>
<td>34.66</td>
<td>40.66</td>
<td>37.42</td>
</tr>
<tr>
<td>Dictionary-based tagger</td>
<td>47.30</td>
<td>47.82</td>
<td>47.56</td>
</tr>
<tr>
<td>TBL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words only</td>
<td>47.08</td>
<td>36.65</td>
<td>41.22</td>
</tr>
<tr>
<td>Dictionary-based tagger</td>
<td>56.80</td>
<td>34.62</td>
<td>43.02</td>
</tr>
<tr>
<td>SVM ($N = 2$, full training set, max. recall)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preceding class labels</td>
<td>69.16</td>
<td>19.74</td>
<td>30.72</td>
</tr>
<tr>
<td>Preceding class labels and part-of-speech</td>
<td>70.18</td>
<td>19.72</td>
<td>30.79</td>
</tr>
<tr>
<td>Preceding class labels and dictionary-based tagger with additional suffix features</td>
<td>65.00</td>
<td>45.41</td>
<td>53.48</td>
</tr>
<tr>
<td>MaxEnt ($N = 1$, Viterbi w/o greedy extraction, max. recall)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W/o dictionary</td>
<td>71.10</td>
<td>42.31</td>
<td>53.05</td>
</tr>
<tr>
<td>With dictionary</td>
<td>73.37</td>
<td>47.76</td>
<td>57.86</td>
</tr>
<tr>
<td>With dictionary, two tags only $(1,0)$</td>
<td>66.41</td>
<td>44.74</td>
<td>53.46</td>
</tr>
<tr>
<td>REX</td>
<td>14.68</td>
<td>31.83</td>
<td>20.09</td>
</tr>
<tr>
<td>Accurate</td>
<td>32.39</td>
<td>45.87</td>
<td>37.97</td>
</tr>
</tbody>
</table>
Another Approach: Learning an NER Model from Labeled Data

- given a corpus of labeled sentences, learn a model to recognize named entities

The focal adhesion kinase is highly expressed in rat osteoclasts in vivo.

NER with a Probabilistic Sequence Model

“Analysis of the FLT-3 protein in progenitor cells at different…”
Features for NER

- in addition to the words themselves, we may want to use other features to characterize the sequence

<table>
<thead>
<tr>
<th>type</th>
<th>example</th>
<th>example matching token</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>word=mitogen?</td>
<td>mitogen</td>
</tr>
<tr>
<td>orthographic</td>
<td>is-alphanumeric?</td>
<td>SH3</td>
</tr>
<tr>
<td></td>
<td>has-dash?</td>
<td>interleukin-1</td>
</tr>
<tr>
<td>shape</td>
<td>AA0</td>
<td>SH3</td>
</tr>
<tr>
<td></td>
<td>A_aaaaa</td>
<td>F-actin</td>
</tr>
<tr>
<td>substring</td>
<td>suffix=ase?</td>
<td>kinase</td>
</tr>
<tr>
<td>lexical</td>
<td>is-amino-acid?</td>
<td>leucine</td>
</tr>
<tr>
<td></td>
<td>is-Greek-letter?</td>
<td>alpha</td>
</tr>
<tr>
<td></td>
<td>is-Roman-numeral?</td>
<td>II</td>
</tr>
<tr>
<td>part-of-speech</td>
<td>is-noun?</td>
<td>membrane</td>
</tr>
</tbody>
</table>

Conditional Random Fields for NER

[Lafferty et al., 2001]

- first-order CRFs define conditional probability of label sequence $y$ given input sequence $o$ to be:

$$P(y | o) = \frac{1}{Z_o} \exp \left( \sum_{i=1}^{L} \sum_{k=1}^{F} \lambda_k f_k (y_{i-1}, y_i, o_i) \right)$$

$\lambda_k$: weight on $k^{th}$ feature
Conditional Random Fields for NER

- the CRF is an undirected graphical model
- the features are used to assess the “compatibility” of the values assigned to each clique

\[
P(y \mid o) = \frac{1}{Z_o} \exp \left( \sum_{i=1}^{L} \sum_{k=1}^{F} \lambda_k f_k (y_{i-1}, y_i, o_i) \right)
\]

- word='analysis' capitalized
- word='of'
- word='flt-3' has-dash
- is-alphanumeric
- next-word='protein'
- word='in'
- word='progenitor' prefix='pro'
- next-word='cells'
Conditional Random Fields for NER

- the NER task involves finding the most probable sequence of labels given the observed sentence

\[ P(y_s | o_s) \]

- this can be done using a variant of the Viterbi algorithm

\[ \sum_s P(y_s | o_s) \]

sum over training sentences
Comparison of NER Systems
NLPBA Workshop (COLING 2004)

Comparison of NER Systems
BioCreative Workshop (BMC Bioinformatics 2005)
The Information Extraction Task: Relation Extraction

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the PRP20 gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected. SRM1 mutants, defective in the same gene, influence the signal transduction pathway for the pheromone response . . . By immunofluorescence microscopy the PRP20 protein was localized in the nucleus. Expression of the RCC1 protein can complement the temperature-sensitive phenotype of PRP20 mutants, demonstrating the functional similarity of the yeast and mammalian proteins

Relation Extraction with OpenDMAP
Hunter et al., BMC Bioinformatics 2008

- OpenDMAP employs hand-coded rules to recognize concepts (entities and relations)
- the rules may include
  - words and phrases
  - part-of-speech categories
  - syntactic dependencies among words
  - semantic categories (recognized from dictionaries, NER systems, etc.)
Relation Extraction with OpenDMAP

• a rule for extracting protein-transport relations

protein-transport:= [transported-entity] translocation ( from {det}? [transport origin] )? ( to {det}? [transport destination] )?;

[ ] arguments of extracted relation
{} part-of-speech categories
? optional elements

"...Bax translocation to mitochondria...."

protein-transport:= [transported-entity] translocation ( from {det}? [transport origin] )? ( to {det}? [transport destination] )?;

[ ] arguments of extracted relation
{} part-of-speech categories
? optional elements
Relation Extraction with OpenDMAP

- some more expressive rules for extracting protein-transport relations

\[
\text{protein-transport} := ( [ \text{transported-entity} ] \text{transportation} ) \\
\quad @ ( \text{from} \{ \text{det}\} \? [ \text{transport-origin} ] ) \\
\quad @ ( \text{to} \{ \text{det}\} \? [ \text{transport-destination} ] ) ;
\]

\[
\text{protein-transport} := ([ \text{transported-entity} \text{dep:x} ] \_ \text{[action action-transport head:x]} ) \\
\quad @ ( \text{from} \{ \text{det}\} \? [ \text{transport-origin} ] ) \\
\quad @ ( \text{to} \{ \text{det}\} \? [ \text{transport-destination} ] ) ;
\]

- arguments of extracted relation
- part-of-speech categories
- optional elements
- 0 or more tokens
- optional elements, occur before/after required phrase
- dependency relationship

A Dependency Parse

- a dependency parse relates each word to other words in the sentence that depend on it

The \text{sigmaB}-dependent promoter drives expression of \text{yvyD} under stress conditions and after glucose starvation whereas a \text{sigmaH}-dependent promoter is responsible for \text{yvyD} transcription.
Relation Extraction with OpenDMAP

"...HIF-1alpha which is present in glomus cells translocates to the nucleus...."

protein-transport:= ([transported-entity dep:x] _
   [action action-transport head:x])
   @ (from {det}? [transport origin])
   @ (to {det}? [transport destination]);

The Event Extraction Task

given
   passages of text
   event types of interest

extract
   a (possibly related) set of events described in the text
The Event Extraction Task

- In 2009 and 2011 there have been “shared tasks” focusing on event extraction
  - Publicly available training corpus with annotated events
  - Server that evaluates predicted events on a test corpus

- Each extracted event consists of
  - A trigger: a word or phrase indicating a specific relation
  - One or more arguments: each of which is an entity or another event

---

Event vocabulary for the *BioNLP Shared Task* corpus

<table>
<thead>
<tr>
<th>Event class</th>
<th>Event type</th>
<th>Arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMPLE</td>
<td>Gene expression</td>
<td>Theme(P)</td>
</tr>
<tr>
<td></td>
<td>Transcription</td>
<td>Theme(P)</td>
</tr>
<tr>
<td></td>
<td>Protein catabolism</td>
<td>Theme(P)</td>
</tr>
<tr>
<td></td>
<td>Phosphorylation</td>
<td>Theme(P)</td>
</tr>
<tr>
<td></td>
<td>Localization</td>
<td>Theme(P)</td>
</tr>
<tr>
<td>BINDING</td>
<td>Binding</td>
<td>Theme(P)+</td>
</tr>
<tr>
<td>REGULATION</td>
<td>Regulation</td>
<td>Theme(P/E), Cause(P/E)</td>
</tr>
<tr>
<td></td>
<td>Positive Regulation</td>
<td>Theme(P/E), Cause(P/E)</td>
</tr>
<tr>
<td></td>
<td>Negative Regulation</td>
<td>Theme(P/E), Cause(P/E)</td>
</tr>
</tbody>
</table>
SQ22536 suppressed \textbf{gp41}-induced \textbf{IL-10} production

Negative regulation
\hspace{1cm} trigger
\hspace{1cm} theme
\hspace{1cm} suppressed
Positive regulation
\hspace{1cm} cause
\hspace{1cm} trigger
\hspace{1cm} induced
Gene expression
\hspace{1cm} trigger
\hspace{1cm} theme
production
\hspace{1cm} \textbf{IL-10}

Phosphorylation
\hspace{1cm} trigger
\hspace{1cm} theme
 Localization
\hspace{1cm} trigger
\hspace{1cm} theme

…decreased tyrosine phosphorylation and nuclear translocation of \textbf{STAT6}…
A pipeline approach to event extraction

Step 1: recognize triggers

SQ22536 suppressed **gp41** induced **IL-10** production

trigger classifier

Neg_reg Pos_reg Gene_expr

A pipeline approach to event extraction

Step 2: assign Theme arguments

SQ22536 suppressed **gp41** induced **IL-10** production

Neg_reg Pos_reg Gene_expr
A pipeline approach to event extraction

Step 2: assign Theme arguments

SQ22536 suppressed gp41 induced IL-10 production

Neg_reg Pos_reg Gene_expr

A pipeline approach to event extraction

Step 3: assign Cause arguments

SQ22536 suppressed gp41 induced IL-10 production

Neg_reg Pos_reg Gene_expr
A pipeline approach to event extraction

Step 3: assign Cause arguments

SQ22536 suppressed gp41 induced IL-10 production

Neg_reg Pos_reg Gene_expr

A pipeline approach to event extraction

Step 4: construct events

SQ22536 suppressed gp41 induced IL-10 production

Negative regulation

Positive regulation

Gene expression

trigger suppressed

cause gp41

trigger induced

trigger production

theme

theme

theme

cause

trigger

theme

trigger

theme
Classifiers for event extraction
[Vlachos & Craven, CoNLL ‘11]

- dependency paths
- POS tags
- types of the candidate arguments (protein or event?)
- lemmatized words
- etc.

Features for the classifiers are based on

Local decisions are made using classifiers trained with the Passive-Aggressive algorithm [Crammer et al., JMLR 2006]

Learning the classifiers jointly
[Vlachos & Craven, CoNLL ‘11]

- The labeled training corpus enables these classifiers to be trained independently
- We train them jointly using an approach called Searn (Daume et al.)
Event accuracy
Searn vs. standard pipeline

development set
(used like a test set here)

Recall | Precision | F1
---|---|---
standard pipeline: 36.1 | 50.8 | 47.5
Searn pipeline: 69.2 | 59.5 | 54.8

Recall | Precision | F1
---|---|---
test set
standard pipeline: 30.8 | 44.5 | 42.2
Searn pipeline: 67.4 | 59.1 | 50.8

Event accuracy
Searn vs. MLNs

Recall | Precision | F1
---|---|---
MLN (Riedel et al.): 36.9 | 43.7 | 44.5
MLN (Poon & Vanderwende): 55.6 | 58.6 | 59.1
Searn pipeline: 44.4 | 50.8 | 50.8