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

The Named Entity Recognition Task

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(PRP20, nucleus)

The Relation Extraction Task

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

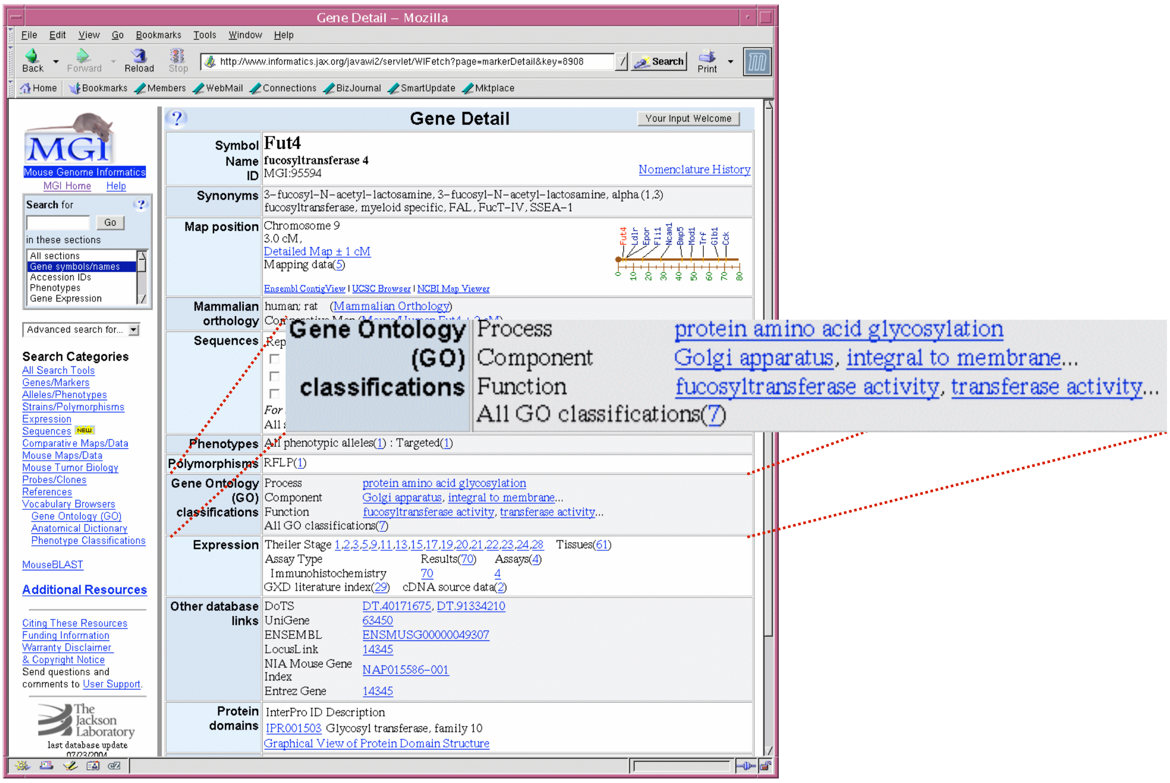
→ subcellular-localization(PRP20, nucleus)

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



Gene Detail

Symbol Fut4
Name fucosyltransferase 4
ID MGI:95594
Nomenclature History

Synonyms 3–fucosyl–N–acetyl–lactosamine, 3–fucosyl–N–acetyl–lactosamine, alpha (1,3) fucosyltransferase, myeloid specific, FAL, FucT–IV, SSEA–1

Map position Chromosome 9
3.0 cM,
Detailed Map ± 1 cM
Mapping data(5)
Ensembl ContigView | UCSC Browser | NCBI Map Viewer

Mammalian orthology human; rat (Mammalian Orthology)
Co

Sequences Rep
□
□
□
For
All

Phenotypes All phenotypic alleles(1) : Targeted(1)

Polymorphisms RFLP(1)

Gene Ontology (GO) classifications
Process protein amino acid glycosylation
Component Golgi apparatus, integral to membrane...
Function fucosyltransferase activity, transferase activity...
All GO classifications(7)

Gene Ontology (GO) classifications
Process protein amino acid glycosylation
Component Golgi apparatus, integral to membrane...
Function fucosyltransferase activity, transferase activity...
All GO classifications(7)

Expression Theiler Stage 1,2,3,5,9,11,13,15,17,19,20,21,22,23,24,28 Tissues(61)

Assay Type	Results(70)	Assays(4)
Immunohistochemistry	70	4

GXD literature index(29) cDNA source data(2)

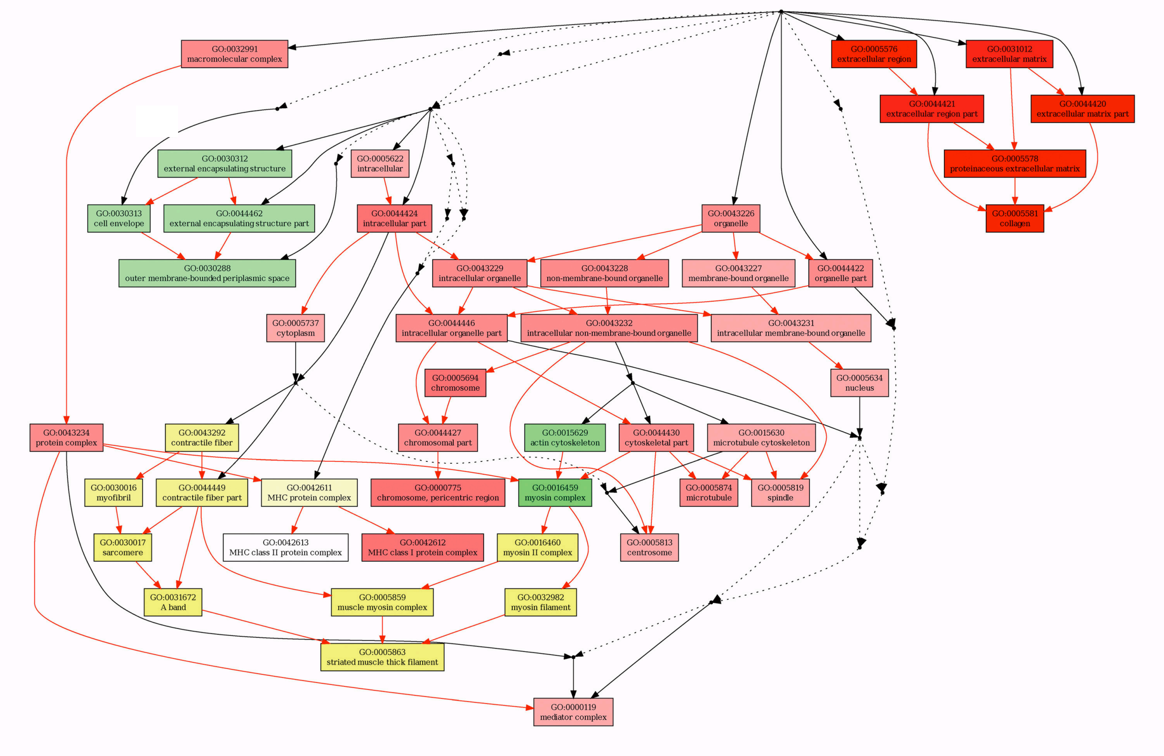
Other database links

DoTS	DT.40171675, DT.91334210
UniGene	63450
ENSEMBL	ENSMUSG00000049307
LocusLink	14345
NIA Mouse Gene Index	NAP015586–001
Entrez Gene	14345

Protein domains
InterPro ID Description
IPR001503 Glycosyl transferase, family 10
Graphical View of Protein Domain Structure

The Gene Ontology

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


GO:0032991 macromolecular complex
GO:0005576 extracellular region
GO:0031012 extracellular matrix
GO:0044421 extracellular region part
GO:0044420 extracellular matrix part
GO:0030312 external encapsulating structure
GO:0005622 intracellular
GO:0005578 proteinaceous extracellular matrix
GO:0030313 cell envelope
GO:0044462 external encapsulating structure part
GO:0044424 intracellular part
GO:0043226 organelle
GO:0005581 collagen
GO:0030288 outer membrane-bounded periplasmic space
GO:0043229 intracellular organelle
GO:0043228 non-membrane-bound organelle
GO:0043227 membrane-bound organelle
GO:0044422 organelle part
GO:0005737 cytoplasm
GO:0044446 intracellular organelle part
GO:0043232 intracellular non-membrane-bound organelle
GO:0043231 intracellular membrane-bound organelle
GO:0005694 chromosome
GO:0005634 nucleus
GO:0043234 protein complex
GO:0043292 contractile fiber
GO:0044427 chromosomal part
GO:0015629 actin cytoskeleton
GO:0044430 cytoskeletal part
GO:0015630 microtubule cytoskeleton
GO:0030016 myofibril
GO:0044449 contractile fiber part
GO:0042611 MHC protein complex
GO:0000775 chromosome, pericentric region
GO:0016459 myosin complex
GO:0005874 microtubule
GO:0005819 spindle
GO:0030017 sarcomere
GO:0042613 MHC class II protein complex
GO:0042612 MHC class I protein complex
GO:0016460 myosin II complex
GO:0005813 centrosome
GO:0031672 A band
GO:0005859 muscle myosin complex
GO:0032982 myosin filament
GO:0005863 striated muscle thick filament
GO:0000119 mediator complex

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

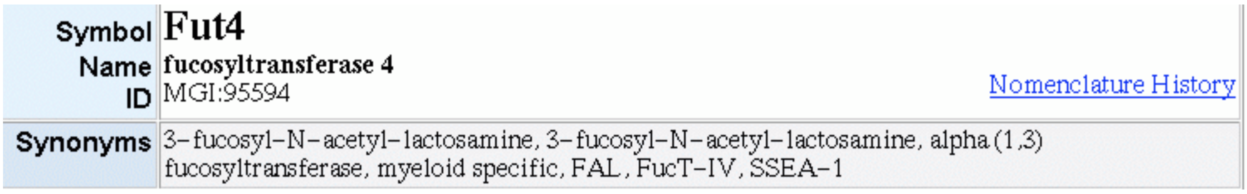
  NF-kappaB
  NF KappaB
  NF-kappa B
  (NF)-kappaB

- 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



Symbol	**Fut4**	
Name	**fucosyltransferase 4**	
ID	MGI:95594	Nomenclature History
Synonyms	3-fucosyl-N-acetyl-lactosamine, 3-fucosyl-N-acetyl-lactosamine, alpha (1,3) fucosyltransferase, myeloid specific, FAL, FucT-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

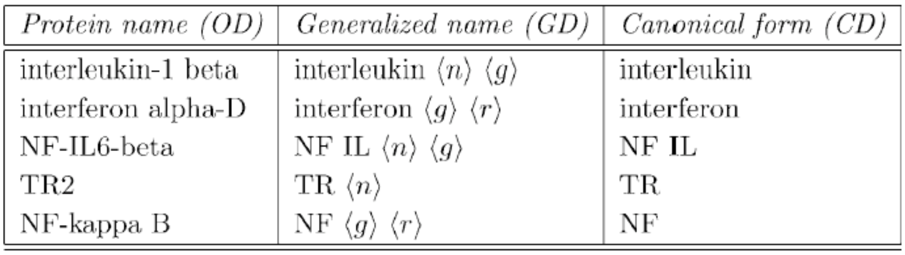
Recognizing Protein Names: Lexical/Syntactic Analysis in Fukuda Approach

- 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]



Protein name (OD)	*Generalized name (GD)*	*Canonical form (CD)*
interleukin-1 beta	interleukin ⟨n⟩ ⟨g⟩	interleukin
interferon alpha-D	interferon ⟨g⟩ ⟨r⟩	interferon
NF-IL6-beta	NF IL ⟨n⟩ ⟨g⟩	NF IL
TR2	TR ⟨n⟩	TR
NF-kappa B	NF ⟨g⟩ ⟨r⟩	NF

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

NER Results from Bunescu et al.

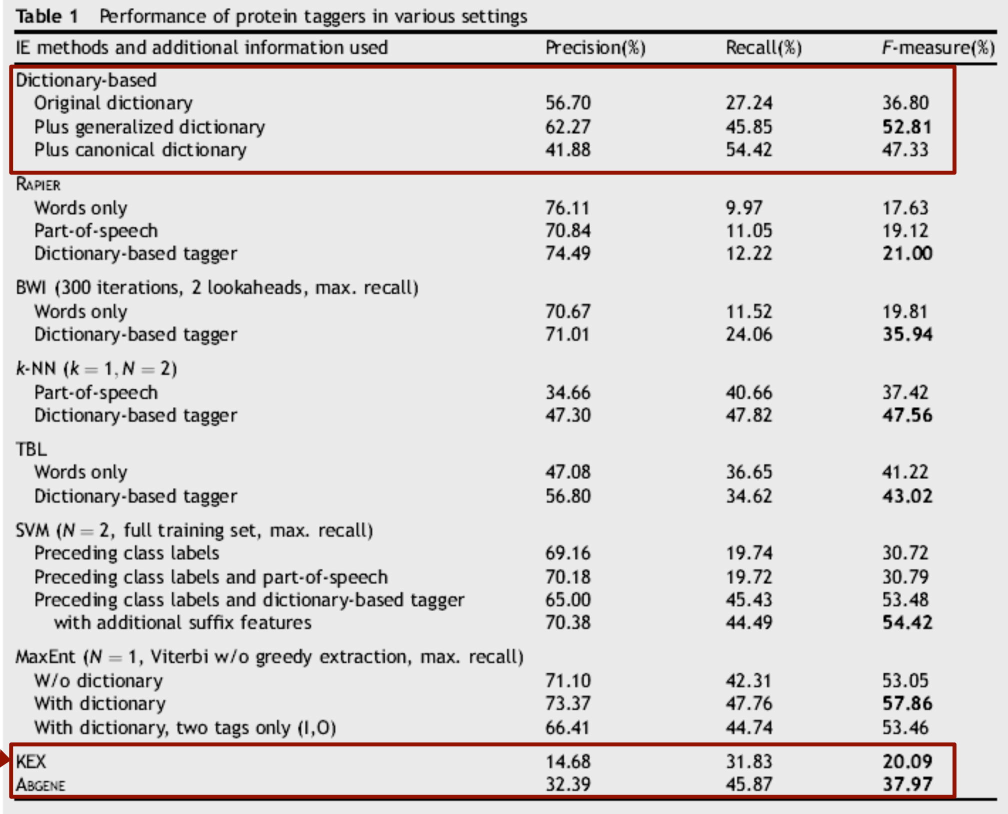


Table 1 Performance of protein taggers in various settings

IE methods and additional information used	Precision(%)	Recall(%)	*F*-measure(%)
Dictionary-based			
Original dictionary	56.70	27.24	36.80
Plus generalized dictionary	62.27	45.85	**52.81**
Plus canonical dictionary	41.88	54.42	47.33
Rapier			
Words only	76.11	9.97	17.63
Part-of-speech	70.84	11.05	19.12
Dictionary-based tagger	74.49	12.22	**21.00**
BWI (300 iterations, 2 lookaheads, max. recall)			
Words only	70.67	11.52	19.81
Dictionary-based tagger	71.01	24.06	**35.94**
k-NN ($k = 1, N = 2$)			
Part-of-speech	34.66	40.66	37.42
Dictionary-based tagger	47.30	47.82	**47.56**
TBL			
Words only	47.08	36.65	41.22
Dictionary-based tagger	56.80	34.62	**43.02**
SVM ($N = 2$, full training set, max. recall)			
Preceding class labels	69.16	19.74	30.72
Preceding class labels and part-of-speech	70.18	19.72	30.79
Preceding class labels and dictionary-based tagger	65.00	45.43	53.48
with additional suffix features	70.38	44.49	**54.42**
MaxEnt ($N = 1$, Viterbi w/o greedy extraction, max. recall)			
W/o dictionary	71.10	42.31	53.05
With dictionary	73.37	47.76	**57.86**
With dictionary, two tags only (I,O)	66.41	44.74	53.46
KEX	14.68	31.83	**20.09**
Abgene	32.39	45.87	**37.97**

Fukuda et al.

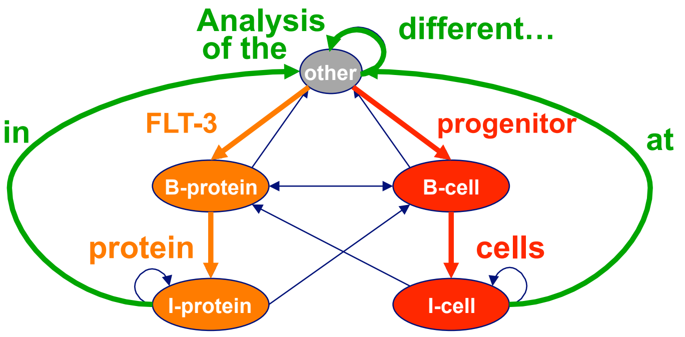
Another Approach: Learning an NER Model from Labeled Data

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

other B-protein I-protein I-protein other other other other B-cell I-cell other other

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

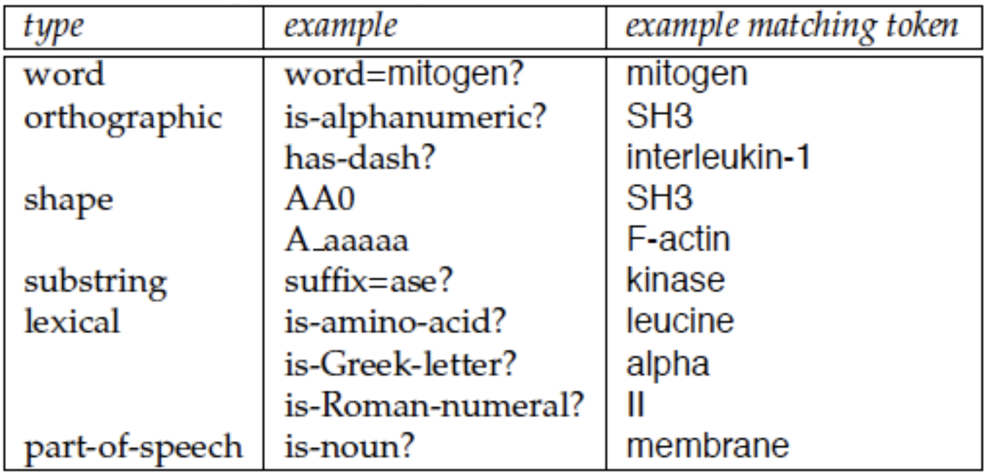
NER with a Probabilistic Sequence Model



"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



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

Conditional Random Fields for NER

[Lafferty et al., 2001]

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

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

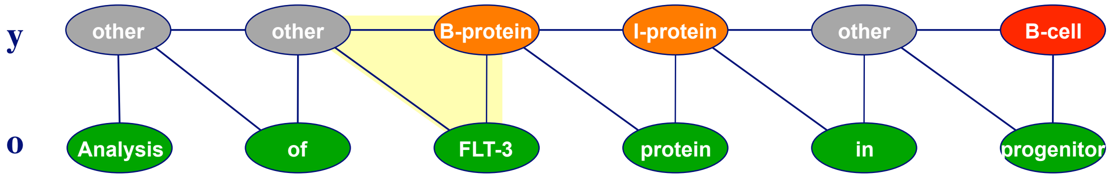
weight on kth feature

kth feature

Conditional Random Fields for NER

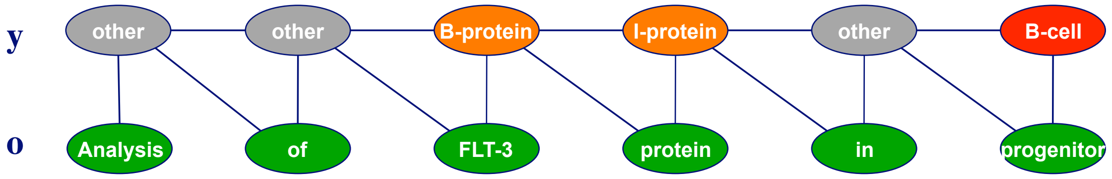
- the CRF is an undirected graphical model
- the features are used to assess the “compatibility” of the values assigned to each clique

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



Conditional Random Fields for NER

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



word=‘analysis’
capitalized

word=‘of’

word=‘flt-3’
has-dash
is-alphanumeric
next-word=’protein’

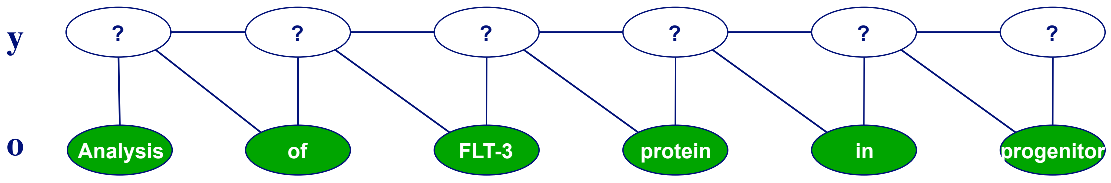
word=‘protein’

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



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

Conditional Random Fields for NER

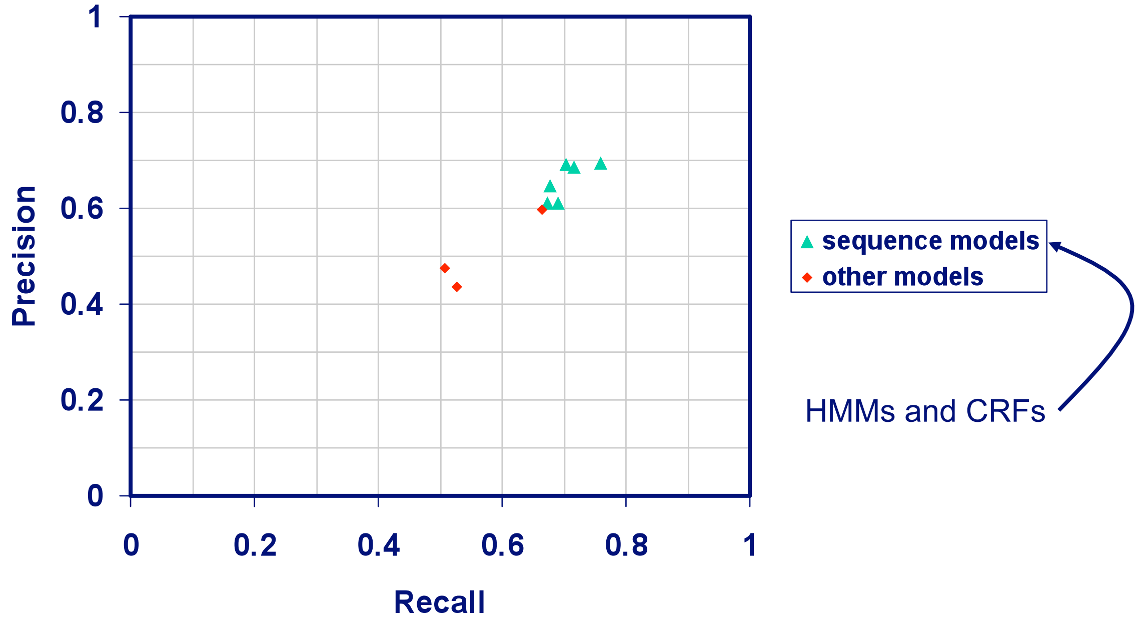
- the training task involves maximizing:

$$\sum_{s} P(\mathbf{y}_s \mid \mathbf{o}_s)$$

sum over
training sentences

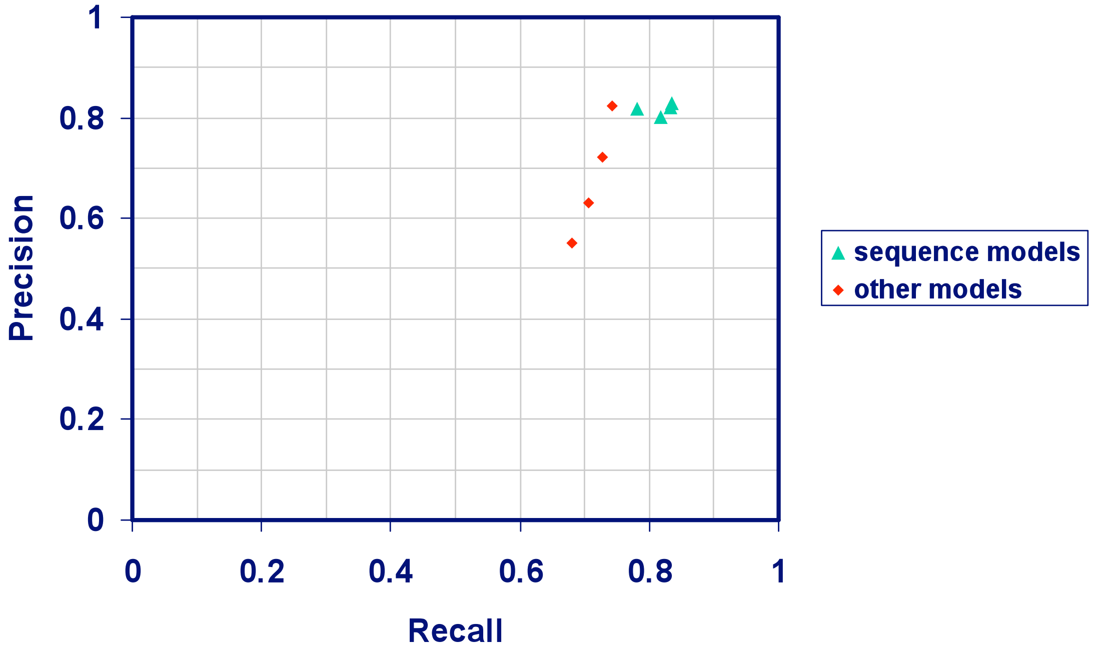
Comparison of NER Systems

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

→ subcellular-localization(PRP20, nucleus)

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 dicitionaries, NER systems, etc.)

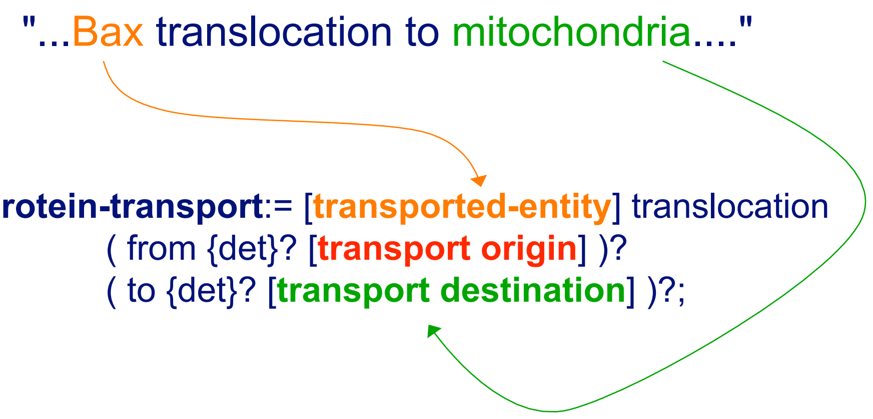
Relation Extraction with OpenDMAP

- 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

Relation Extraction with OpenDMAP



"...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

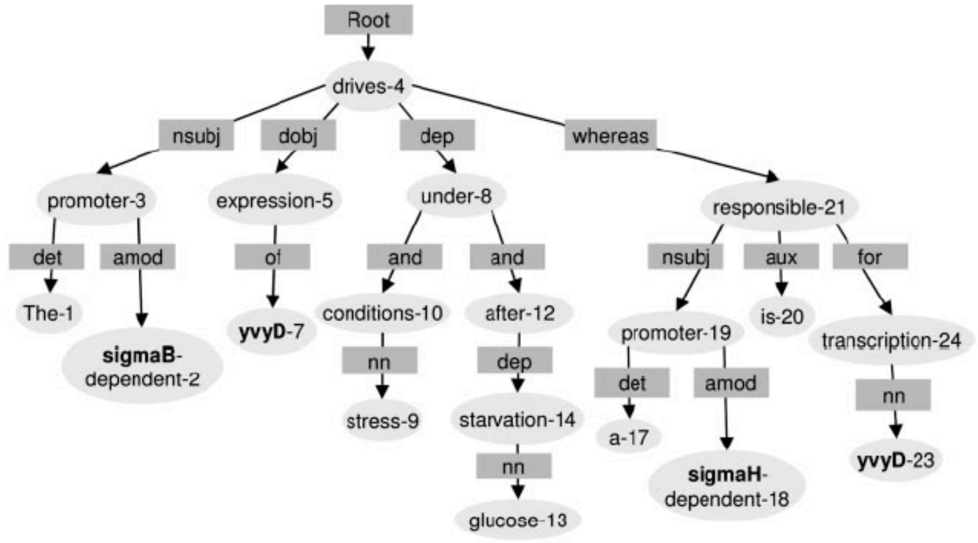
```
protein-transport :=( [ transported-entity ] translocation )
                     @ ( from {det}? [ transport-origin ] )
                     @ ( to {det}? [ transport-destination ] );

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

[ ]	arguments of extracted relation
{ }	part-of-speech categories
?	optional elements
_	0 or more tokens
@	optional elements, occur before/after required phrase
dep:*x*, head:*x*	dependency relationship

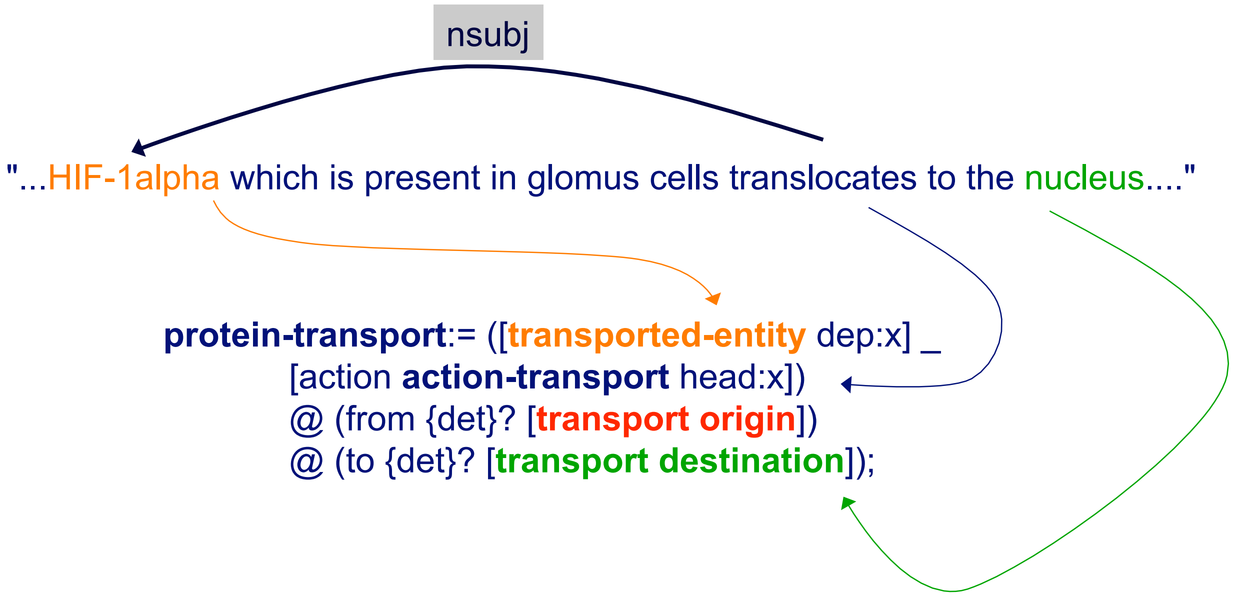
A Dependency Parse

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



The **sigmaB**-dependent promoter drives expression of **yvyD** under stress conditions and after glucose starvation whereas a **sigmaH**-dependent promoter is responsible for **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

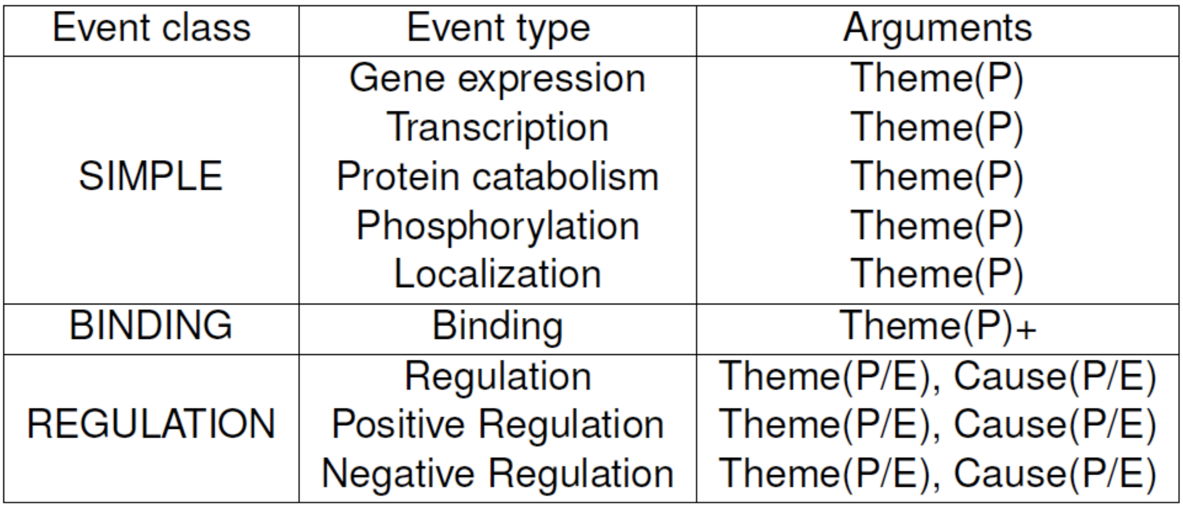
a (possibly related) set of events described in the text

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

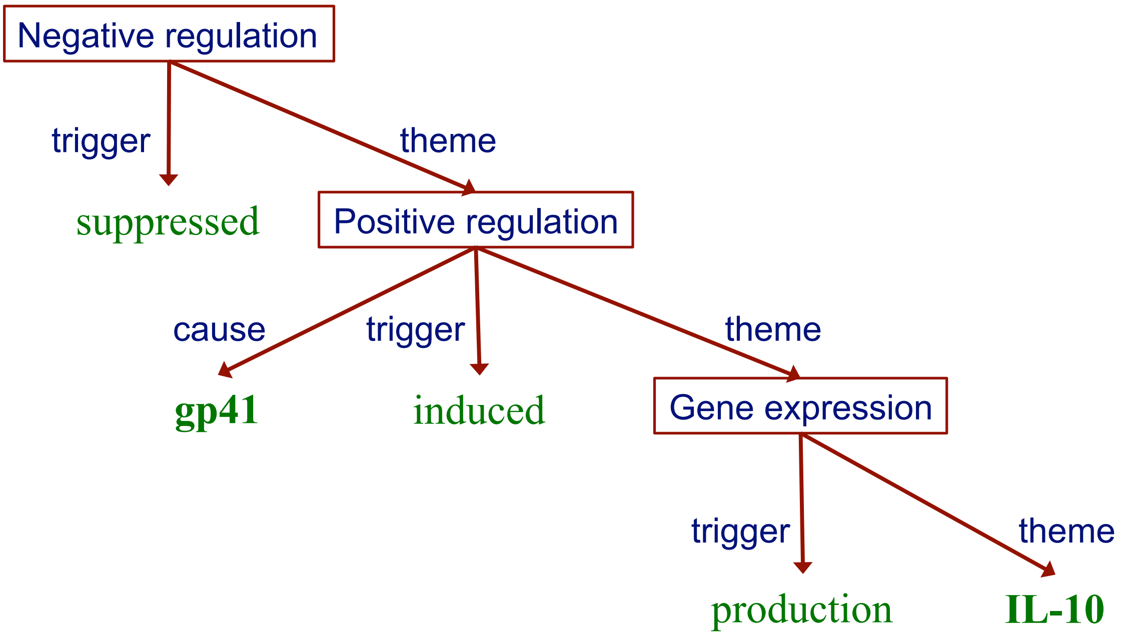
Event vocabulary for the *BioNLP Shared Task* corpus



Event class	Event type	Arguments
SIMPLE	Gene expression	Theme(P)
	Transcription	Theme(P)
	Protein catabolism	Theme(P)
	Phosphorylation	Theme(P)
	Localization	Theme(P)
BINDING	Binding	Theme(P)+
REGULATION	Regulation	Theme(P/E), Cause(P/E)
	Positive Regulation	Theme(P/E), Cause(P/E)
	Negative Regulation	Theme(P/E), Cause(P/E)

Event extraction example

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



Negative regulation

trigger

theme

suppressed

Positive regulation

cause

trigger

theme

gp41

induced

Gene expression

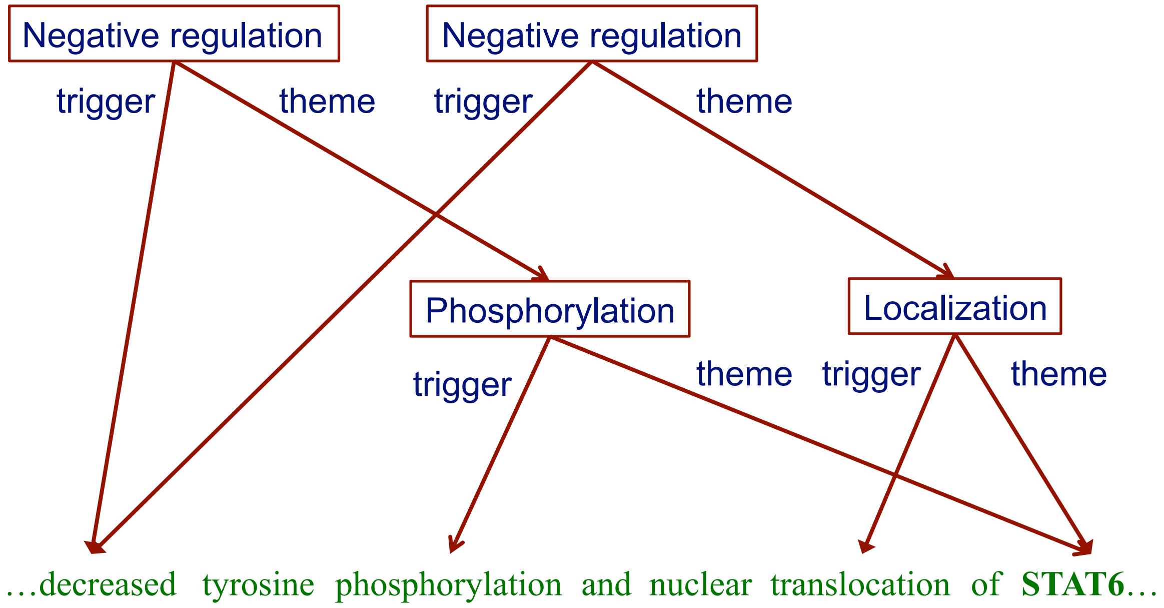
trigger

theme

production

IL-10

Event extraction example



Negative regulation

trigger

theme

Negative regulation

trigger

theme

Phosphorylation

trigger

theme

Localization

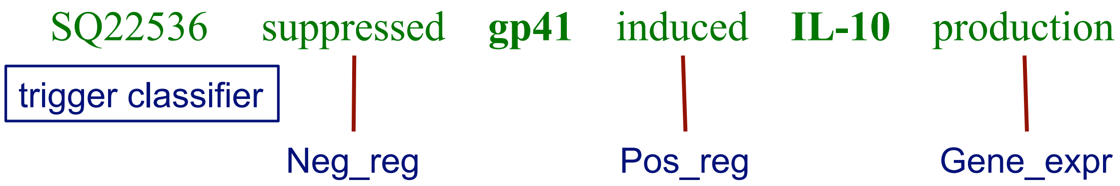
trigger

theme

…decreased tyrosine phosphorylation and nuclear translocation of **STAT6**…

A pipeline approach to event extraction

Step 1: recognize triggers



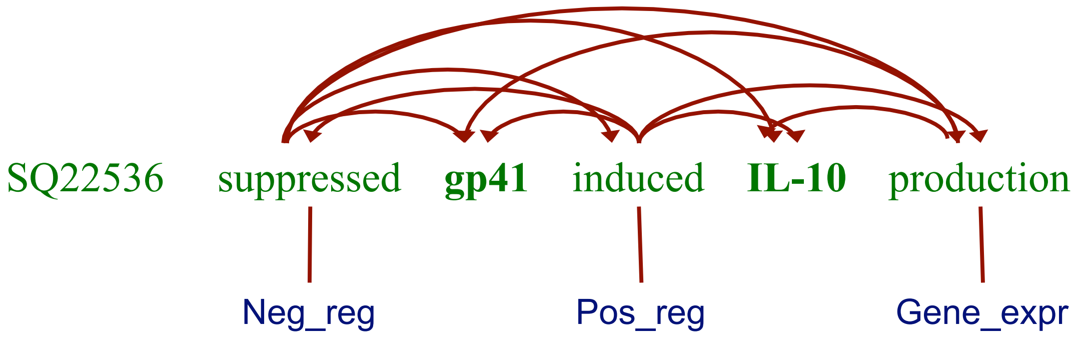
SQ22536 suppressed **gp41** induced **IL-10** production

trigger classifier

Neg_reg Pos_reg Gene_expr

A pipeline approach to event extraction

Step 2: assign Theme arguments

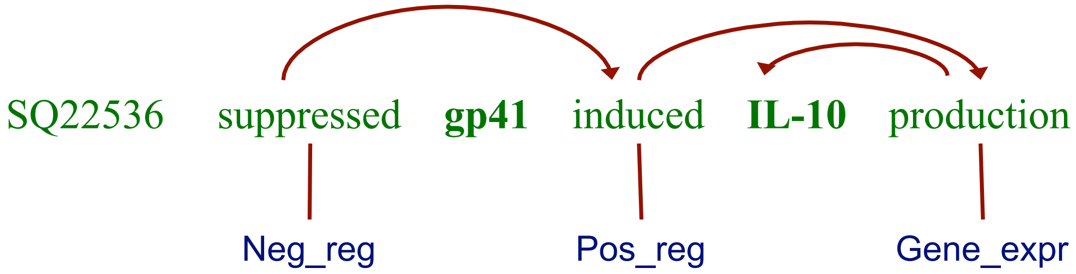


SQ22536 suppressed **gp41** induced **IL-10** production

Neg_reg Pos_reg Gene_expr

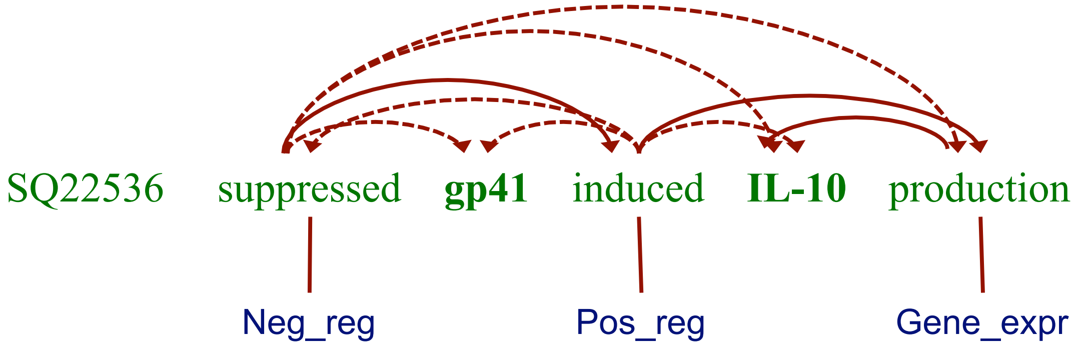
A pipeline approach to event extraction

Step 2: assign Theme arguments



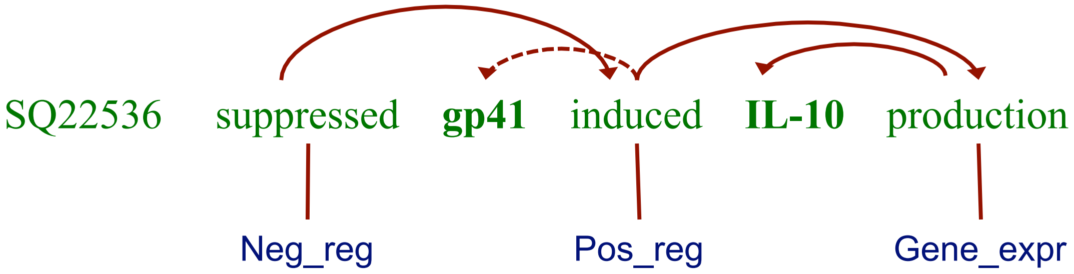
A pipeline approach to event extraction

Step 3: assign Cause arguments



A pipeline approach to event extraction

Step 3: assign Cause arguments



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

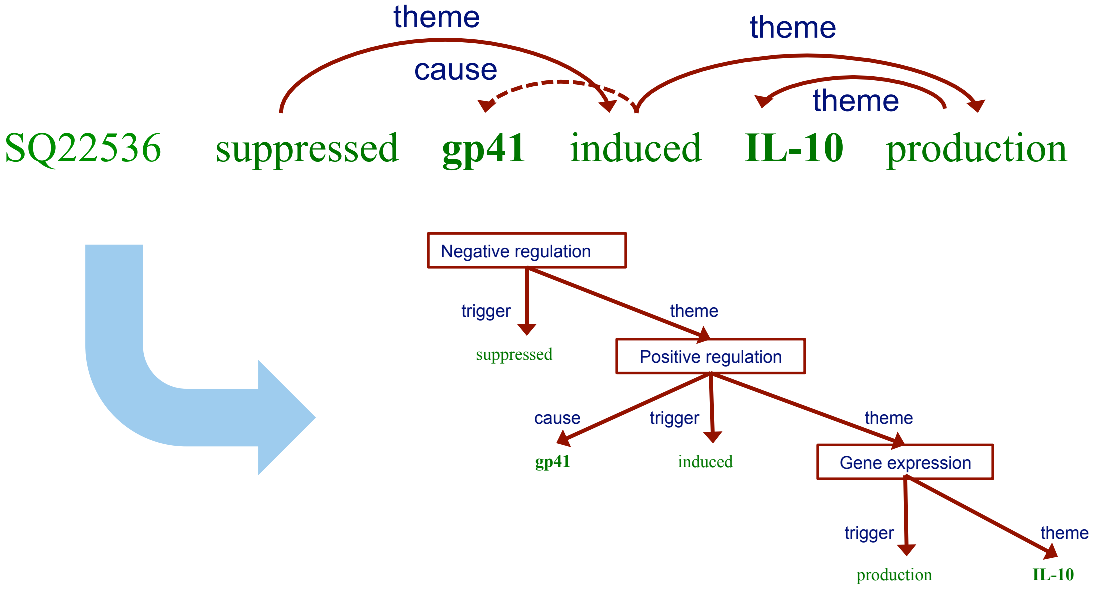
Neg_reg

Pos_reg

Gene_expr

A pipeline approach to event extraction

Step 4: construct events



theme

cause

theme

theme

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

Negative regulation

trigger

theme

suppressed

Positive regulation

cause

trigger

theme

gp41

induced

Gene expression

trigger

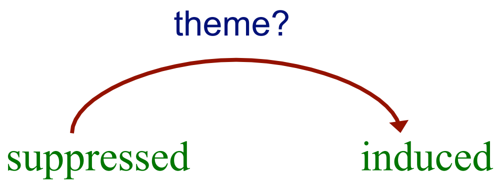
theme

production

IL-10

Classifiers for event extraction

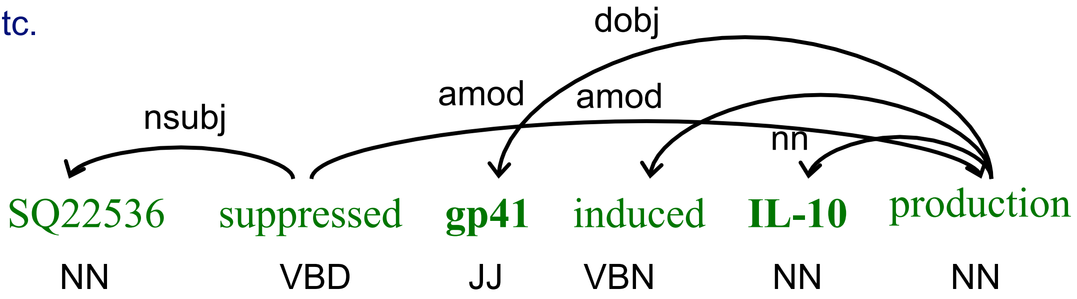
[Vlachos & Craven, CoNLL ‘11]



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

features for the classifiers are based on

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



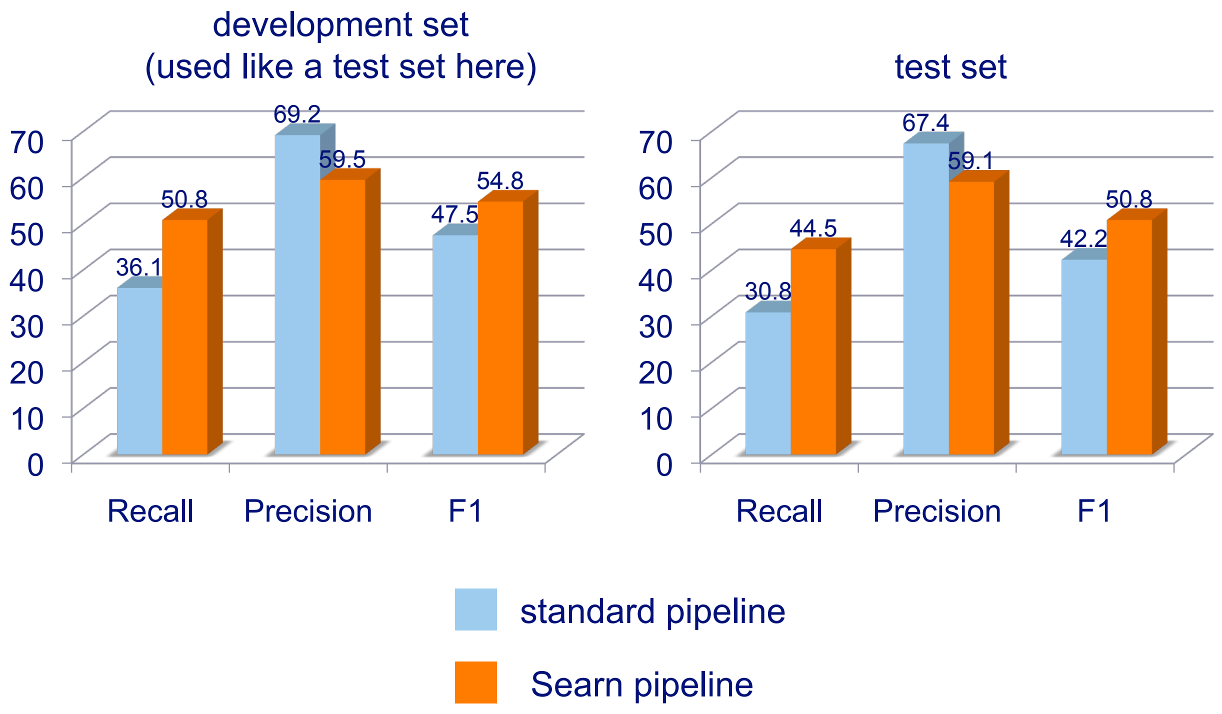
Learning the classifiers jointly

[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.)

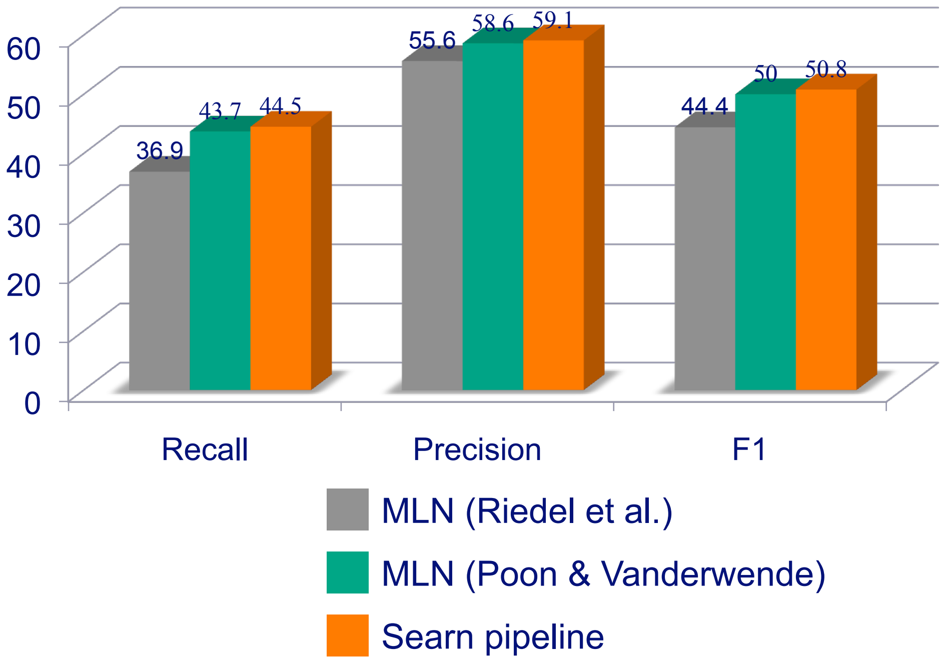
Event accuracy
Searn vs. standard pipeline



development set
(used like a test set here)

test set

Event accuracy
Searn vs. MLNs



Recall
Precision
F1
MLN (Riedel et al.)
MLN (Poon & Vanderwende)
Searn pipeline