Information Extraction from Biomedical Text

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Mark Craven
craven@biostat.wisc.edu
February 2008

Some Important Text-Mining Problems

- hypothesis generation
  Given: biomedical objects/classes of interest (e.g. dieases & dietary factors)
  Do: identify interesting, implied relationships among the objects
- experiment annotation
  Given: a set of genes/proteins exhibiting common behavior in an experiment
  Do: identify commonalities among genes/proteins in the set
- information extraction
  Given: classes, relations of interest
  Do: recognize and extract instances of the classes and relations from documents

Some Important Text-Mining Problems

- document classification
  Given: defined classes of interest
  Do: assign documents to the relevant classes

- ad-hoc retrieval
  Given: a query
  Do: return relevant documents/passages

- improving the accuracy of other inference tasks
  – querying with PSI-BLAST [Chang et al.]
  – predicting subcellular localization of proteins[Hoglund et al.]
  – etc.

The Information Extraction Task: Named Entity Recognition

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 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**)

Motivation for Information Extraction

- motivation for named entity recognition
  - better indexing of biomedical articles
  - assisting in relation extraction
- motivation for relation extraction
  - assisting in the construction and updating of databases
  - providing structured summaries for queries

    What is known about protein X (subcellular & tissue localization, associations with diseases, interactions with drugs, …)?
  - assisting scientific discovery by detecting previously unknown relationships, annotating experimental data

How Do We Get IE Models?

1. encode them by hand

2. learn them from training data

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
  sonic hedgehog

- 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

- these all refer to the same thing
  NF-kappaB
  NF KappaB
  NF-kappa B
  (NF)-kappaB

- there may be references to gene/protein families
  OLE1-4 ⇒
  OLE1   OLE2   OLE3   OLE4

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

Relation Extraction: Representing Sentence Structure in Learned Models

- [Skounakis, Ray & Craven, *IJCAI* '03]
- hidden Markov models (HMMs) have proven to be a good approach for learning IE models
  - can naturally handle relations
  - scale well to long sequences, large data sets
  - provide estimates of uncertainty
  - provide good predictive accuracy in practice
- typically these HMMs have a "flat" structure, and are able to represent relatively little about grammatical structure
- how can we provide HMMs with more information about sentence structure?

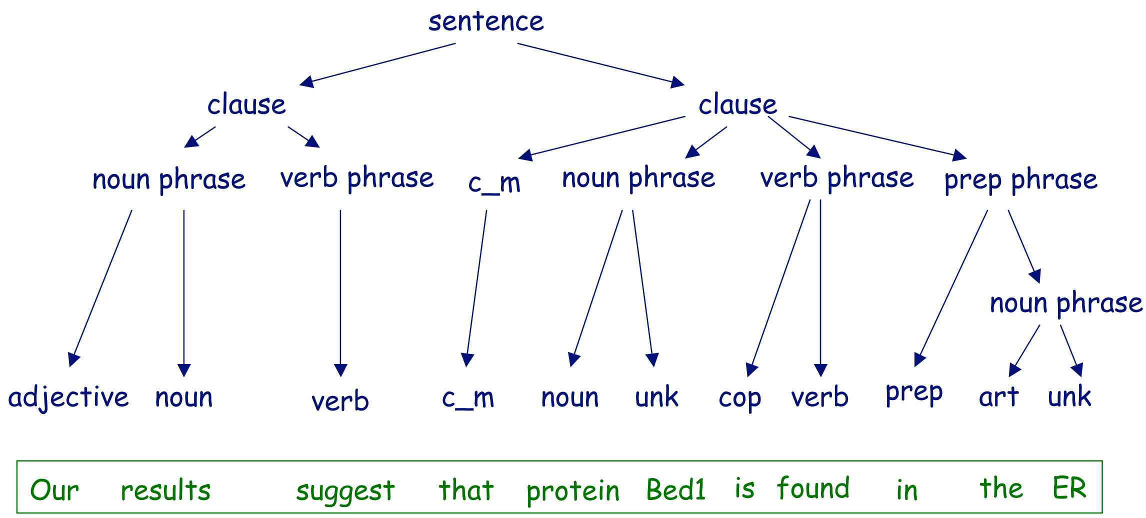
Representing Sentences as Sequences of Tokens

- we can represent sentences as sequences of tokens
- for training sequences we also have labels associated with tokens

	Our
	results
	suggest
	that
	protein
PROTEIN	Bed1
	is
	found
	in
	the
LOCATION	ER

Representing Sentences

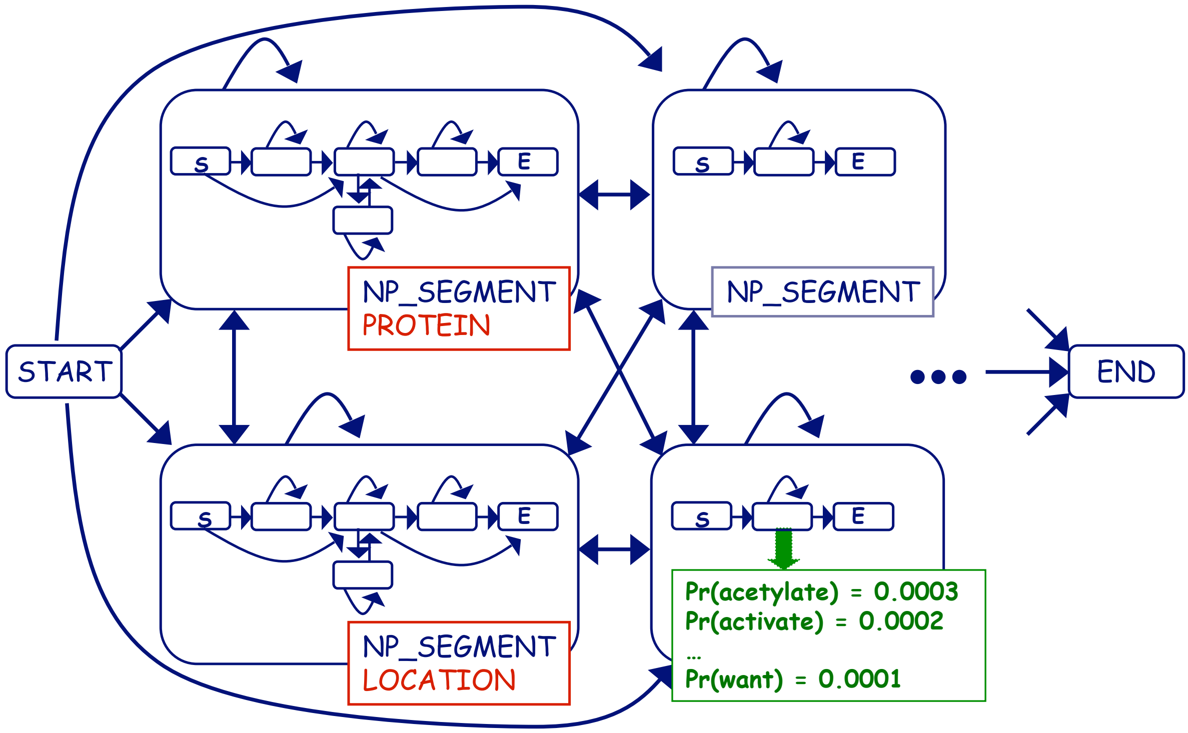
- we first process sentences by analyzing them with a shallow parser (Sundance, [Riloff et al., 98])



Representing Sentences as Nested Sequences of Tokens

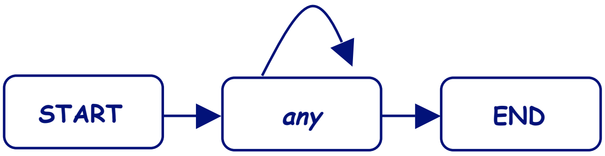
NP_segment	adjective	Our
	noun	results
VP_segment	verb	suggest
c_m	c_m	that
NP_segment: PROTEIN	noun	protein
	unknown: PROTEIN	Bed1
VP_segment	cop	is
	verb	found
prep	prep	in
NP_segment: LOCATION	art	the
	unknown: LOCATION	ER

Hierarchical HMMs for IE

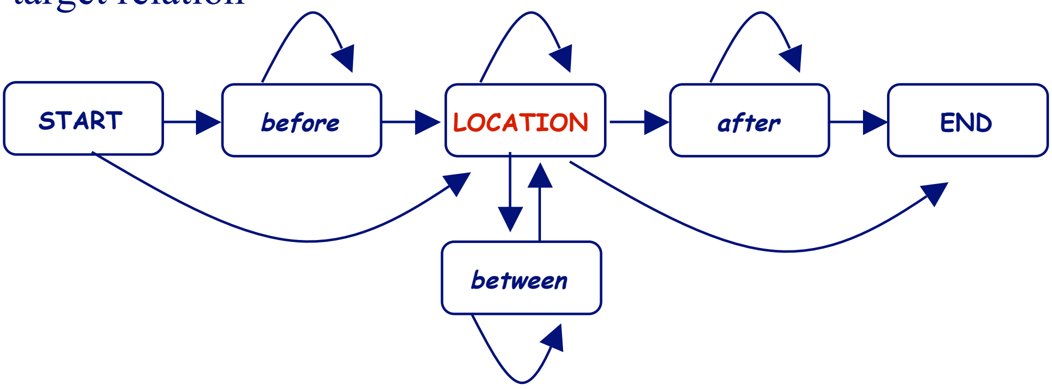


Word-Level HMMs

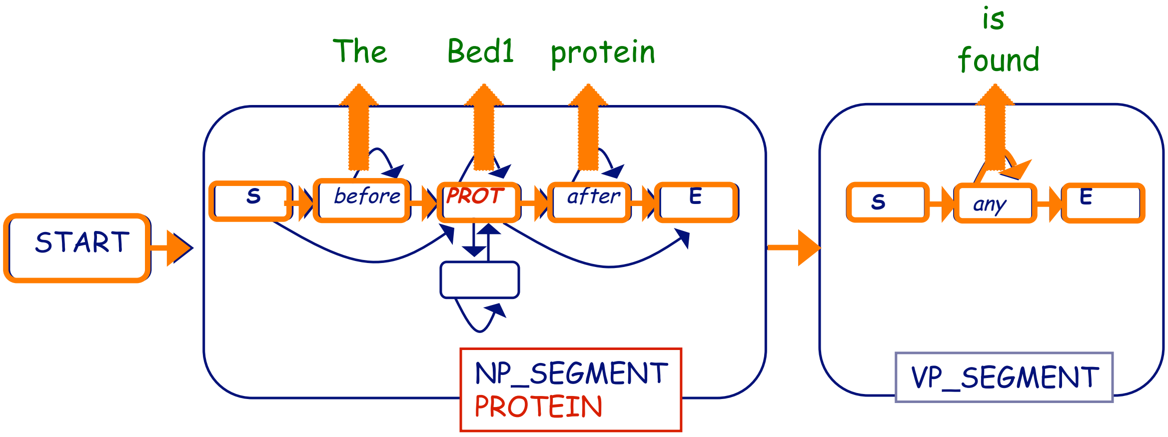
- models for phrase-level states that don't represent a domain of the target relation



- models for phrase-level states that represent <u>one</u> domain of the target relation



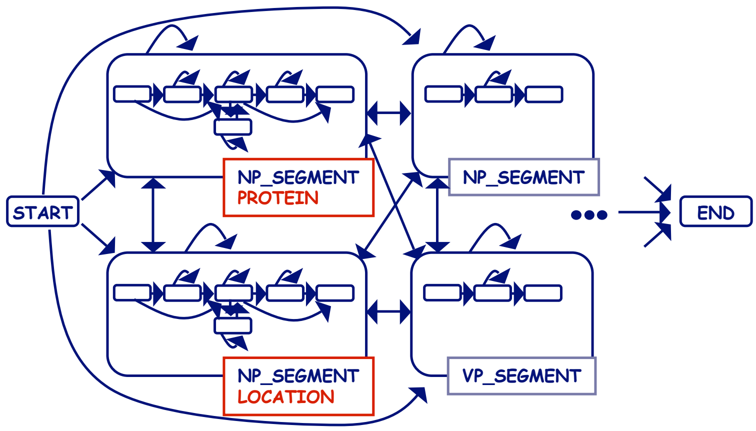
Explaining a Sequence



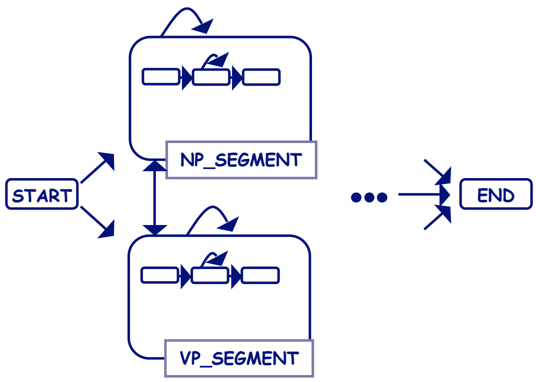
NP_segment	art	The
	unknown: PROTEIN	Bed1
	noun	protein
VP_segment	cop	is
	verb	found

Incorporating a Null Model

- *positive* model trained on sentences labeled with target relations



- *null* model trained on other sentences



Discriminative Training

- In *generative* training, estimate parameters $\hat{\theta}$ such that

$$\hat{\theta} = \arg\max_{\theta} \prod_i \Pr(c_i, s_i \mid \theta)$$

where s_i is the observable sequence and c_i is the sequence of labels for the *i* th instance

- We use a *discriminative* training algorithm [Krogh '94]

$$\hat{\theta} = \arg\max_{\theta} \prod_i \Pr(c_i \mid s_i, \theta)$$

Discriminative Training

Krogh's method provides an on-line, update rule:

$$\theta_j^{new} = N(\theta_j^{old} + \eta(m_j^i - n_j^i))$$

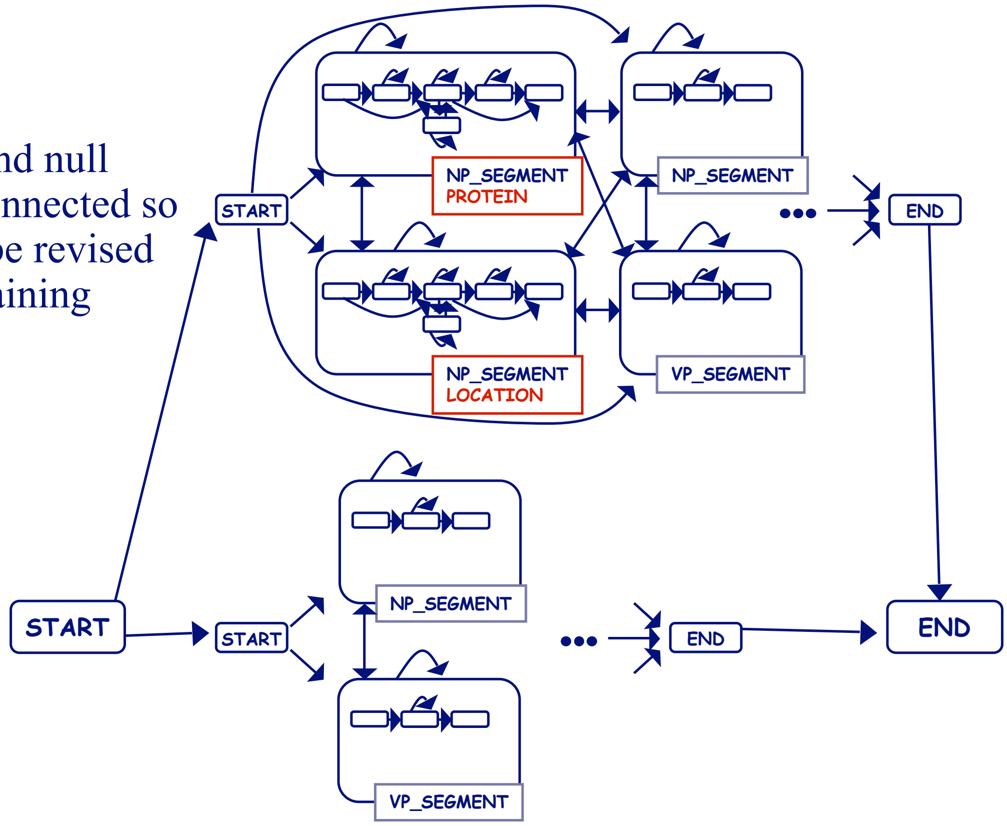
m_j^i : expected number of times θ_j used by *i* th sentence on ***correct*** paths

n_j^i : expected number of times θ_j used by *i* th sentence on ***all*** paths

N : normalizing factor

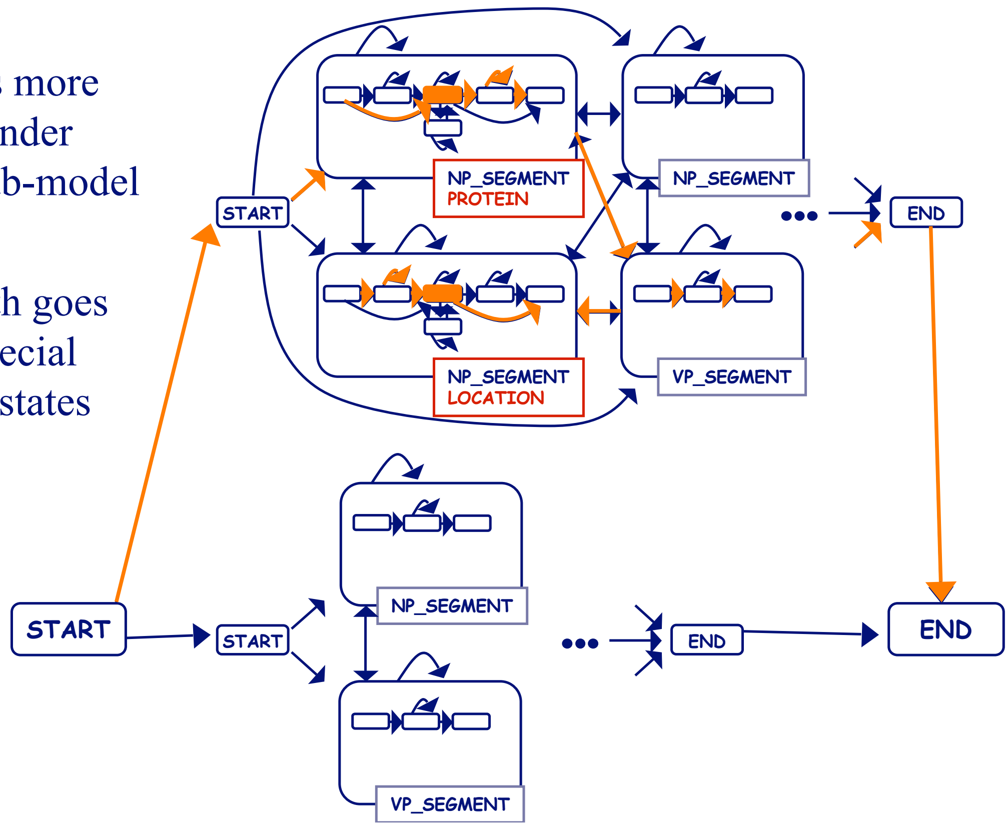
Null Model & Discriminative Training

- positive and null models connected so both can be revised for *any* training instance



Extract a Relation Instance If…

- sentence is more probable under positive sub-model
- Viterbi path goes through special extraction states



Representing More Local Context

- we can have the word-level states represent more about the local context of each emission
- partition sentence into *overlapping trigrams*

"... the/ART Bed1/UNK protein/N is/COP located/V ..."

$w_{-1}, p_{-1} \quad w_0, p_0 \quad w_{+1}, p_{+1}$

$w_{-1}, p_{-1} \quad w_0, p_0 \quad w_{+1}, p_{+1}$

$w_{-1}, p_{-1} \quad w_0, p_0 \quad w_{+1}, p_{+1}$

Representing More Local Context

- states emit trigrams with probability: $t = \langle w_{-1}, w_0, w_{+1}, p_{-1}, p_0, p_{+1} \rangle$

$$\Pr(t) = \Pr(w_{-1})\Pr(w_0)\Pr(w_{+1})\Pr(p_{-1})\Pr(p_0)\Pr(p_{+1})$$

- note the independence assumption above: we compensate for this naïve assumption by using a *discriminative* training method [Krogh '94] to learn parameters

Experimental Evaluation

- hypothesis: we get more accurate models by using a richer representation of sentence structure in HMMs
- compare predictive accuracy of various types of models/representations
  - hierarchical w/context features
  - hierarchical
  - phrases only
  - tokens w/part of speech
  - tokens only

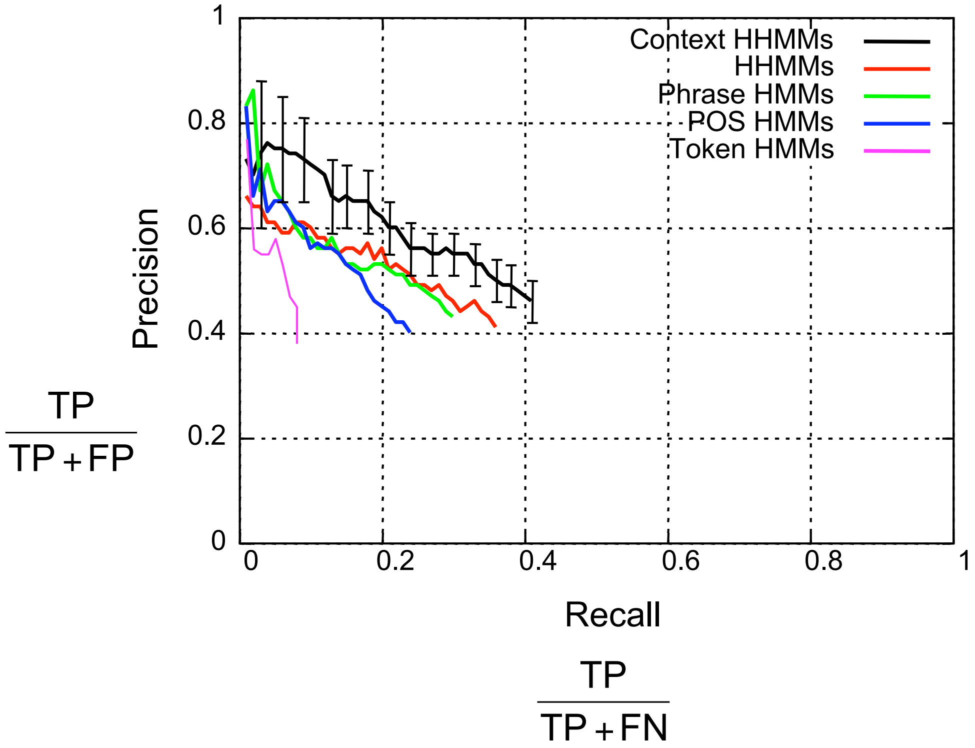
more grammatical information

- 5-fold cross validation on 3 data sets

Data Sets for Learning to Extract Relations

- subcellular_localization(PROTEIN, LOCATION)
  - tuples from YPD database
  - 769 positive, 6193 negative sentences from MEDLINE abstracts
  - 939 tuples (402 distinct)
- disorder_association(GENE, DISEASE)
  - tuples from OMIM database
  - 829 positive, 11685 negative sentences
  - 852 tuples (143 distinct)
- protein_protein_interaction(PROTEIN, PROTEIN)
  - tuples from MIPS database
  - 5446 positive, 41377 negative
  - 8088 tuples (819 distinct)

Extraction Accuracy (protein-location)



Extraction Accuracy (protein-protein interactions)