

Biomedical Text Analysis

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

April 2002

Announcements

- exam
 - 6:00 – 8:00pm, Thursday 5/9
- Thomas Anantharaman guest lecture (next Monday)
- reading for next Wednesday
 - Chapter 9, section 10.1 of Durbin et al.

Some Tasks in Biomedical Text Analysis

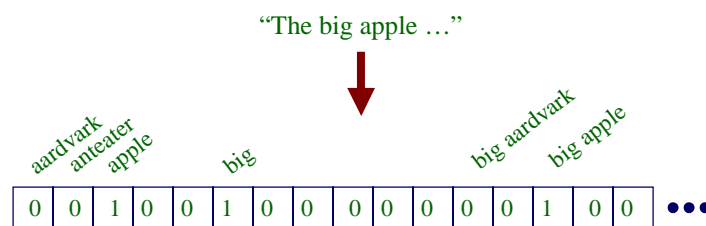
- ✓ extracting keywords/keyphrases for annotating gene/protein families [Andrade & Valencia, 1997; Shatkay et al., 2000]
- recognizing gene/protein names [Fukuda et al., 1998]
- identifying relationships that are implicitly, but not explicitly described in the literature [Swanson & Smalheiser, 1997]
- extracting instances of predefined relations [Craven & Kumlien 1999; Ray & Craven, 2001, etc.]
- using text to improve accuracy of PSI-BLAST queries [Chang et al. 2001]

Task: Automatic Annotation

- Genes, Themes and Microarrays. Shatkay, Edwards, Wilbur & Boguski. *ISMB* 2000
- **given:** a set of genes with a “kernel” document for each
- **return:**
 - top-ranked words in theme for each gene
 - list of most similar genes, in terms of associated documents

Representing Documents

- Shatkay et al. represent documents using fixed-length vectors
 - this is a common approach in many text processing systems (e.g. search engines)
- elements in vector represent occurrences of individual words (unigrams) and pairs of adjacent words (bigrams)



Themes

- a *theme*, T , is a set of documents discussing a common topic
- the occurrence of a given term t_i in a theme document d is represented by

$$p_i^T \equiv \Pr(t_i \in d \mid d \in T)$$

- thus for every term in the vocabulary, we can characterize how likely it is to occur in a document on theme T

Theme Example

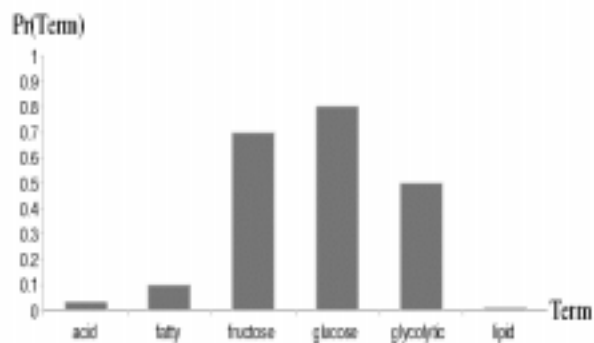


Figure from H. Shatkay et al., *ISMB* 2000

Other Parameters

- Shatkay et al. use similar parameters to represent
 - the occurrence of each term given that document d is not in the theme

$$q_i^T \equiv \Pr(t_i \in d \mid d \notin T)$$

- the occurrence of the term regardless of whether d is on-theme or off-theme

$$DB_i \equiv \Pr(t_i \in d \mid d \in DB)$$

- the prob that a term occurrence is best explained by DB probability or by on-theme/off-theme probabilities

$$\lambda_i$$

Model for “Generating” Documents

- we can think of the document vectors as having been generated from a model with these parameters

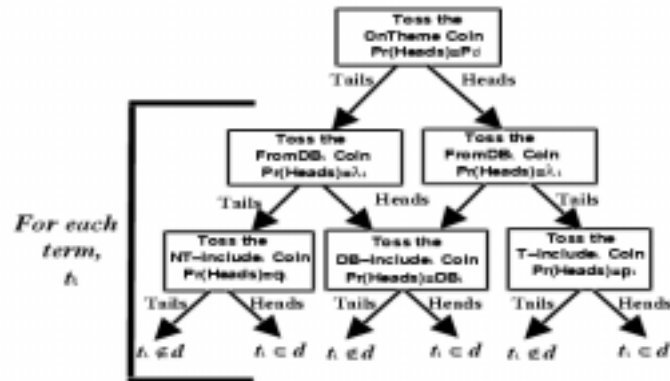


Figure from H. Shatkay et al., *Advances in Digital Libraries* 2000

Finding Themes

- **given:** a DB of documents and a “kernel” document
- **do:**
 - determine the parameters characterizing the theme T
 - determine the documents belonging to T
- if we knew the documents in T , it would be easy to determine the parameters
- if we knew the parameters, it would be easy to determine the documents in T
- but initially, we don’t know either

Finding Themes

- Shatkay et al. solve this problem using EM

E-step: compute likelihood for each document that it's in same theme as kernel

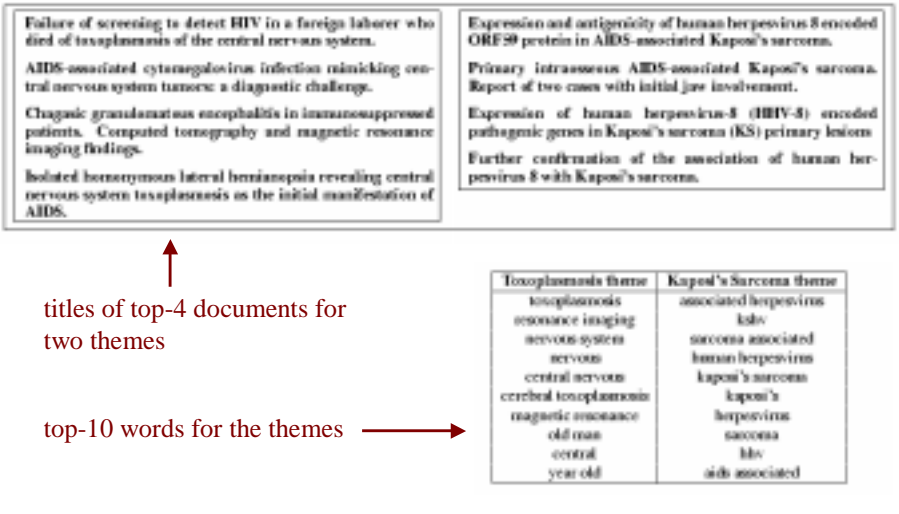
M-step: find new parameters that maximize the likelihood of this partition into theme/off-theme documents

Finding Themes: Output

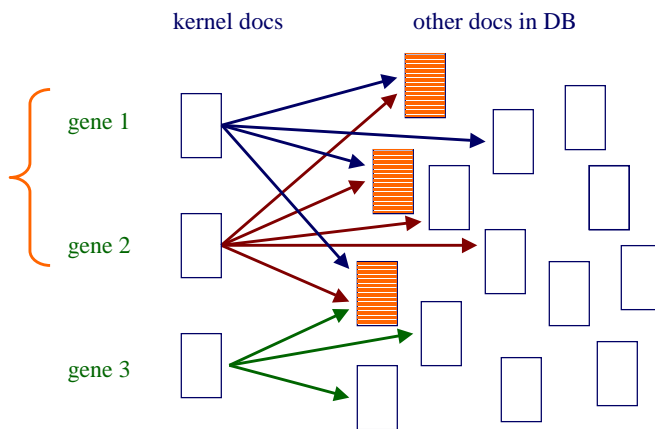
- this EM process is run once for each gene/kernel document
- the results returned for each gene are
 - a list of the most highly weighted ($\frac{p_i^T}{q_i}$) words in the associated theme
 - a list of the most on-theme documents

Finding Themes: Example

- Shatkey et al. have applied this method to find themes in the AIDS literature [*Advances in Digital Libraries*, 2000]



Finding Related Genes

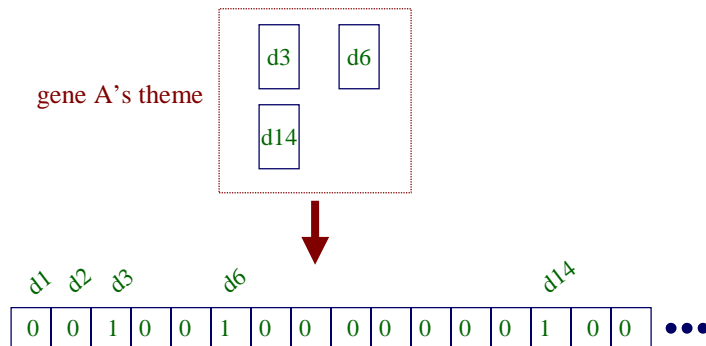


step 1: given kernel documents, find themes

step 2: given themes, find related genes

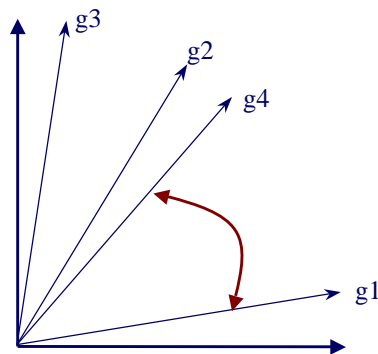
Representing Genes

- represent each gene using fixed-length vector in which each element corresponds to a document
- put a 1 in a given element if the associated document is strongly in the gene's theme



The Vector Space Model

- the similarity between two genes can be assessed by the similarity of their associated vectors
- this is a common method in information retrieval to assess document similarity; here we are assessing gene similarity



Vector Similarity

- one way to determine vector similarity is the cosine measure:

$$\cos(\vec{a}, \vec{b}) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \sqrt{\sum_i b_i^2}}$$

- if the vectors are normalized, we can simply take their dot product

Experiment

- analyzed 408 yeast genes
- documents = abstracts
- kernel documents: oldest reference for each gene in SGD
- database: 33,700 yeast-related documents

Experimental Results

| Kernel (PMID, Gene,Function) | Keywords | Assoc. Genes | Function |
|--|---|--|---|
| 8702485 ELO1 Fatty Acid/ Lipids/ Sterols/ Membranes | fatty acid, fatty, lipids, acid, grown, medium, carbon, synthase, strains, deficient | OLE1 FAA4 FAA3 SUR2 FAA1 ERG2 PSD1 CYB5 PGM1 | (Fatty Acid, Sterol. Met.)* Fatty Acid/Lipids/Sterols/Membranes Fatty Acid/Lipids/Sterols/Membranes Fatty Acid/Lipids/Sterols/Membranes Fatty Acid/Lipids/Sterols/Membranes Fatty Acid/Lipids/Sterols/Membranes Fatty Acid/Lipids/Sterols/Membranes (Fatty Acid, Sterol. Met.)* (Carbohydrates Met.)* |
| 7651133 HXT7 Nutrition | hexose, glucose uptake, glucose conc., fructose, glycolytic, glucose, sugars, uptake, aerobic, utilization | HXT1 RGT2 HXT4 HXT2 GLK1 SEO1 PRB1 AGP1 ZRT1 MIG2 | Nutrition Nutrition Nutrition Nutrition Nutrition (Small Molecules Transport)* (Protein Degradation)* Nutrition Nutrition (Carbohydrates Met.)* |

Figure from H. Shatkay et al., *ISMB* 2000

Announcements

- HW #3 due on Monday
- last year's exam now available on web page
- reading for next week
 - Chapter 9, section 10.1 of Durbin et al.
 - stochastic context free grammars

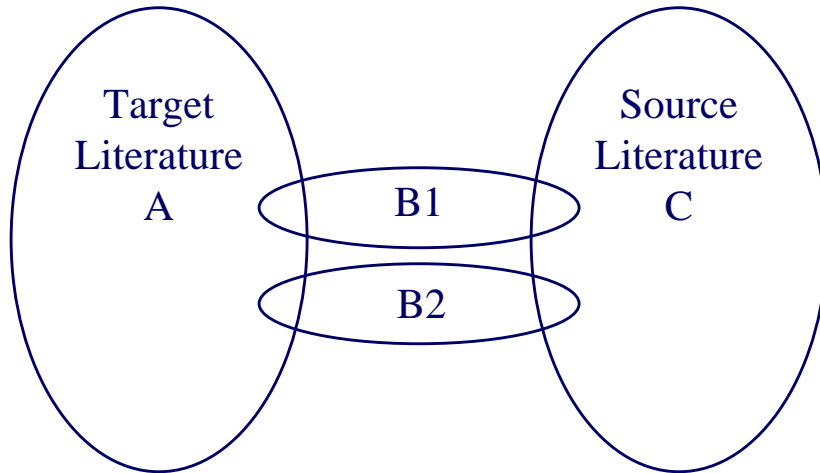
Some Tasks in Biomedical Text Analysis

- extracting keywords/keyphrases for annotating gene/protein families [Andrade & Valencia, 1997; Shatkay et al., 2000]
- ✓ identifying relationships that are implicitly, but not explicitly described in the literature [Swanson & Smalheiser, 1997]
- ✓ recognizing gene/protein names [Fukuda et al., 1998]
- ✓ extracting instances of predefined relations [Craven & Kumlien 1999; Ray & Craven, 2001, etc.]
- using text to improve accuracy of PSI-BLAST queries [Chang et al. 2001]

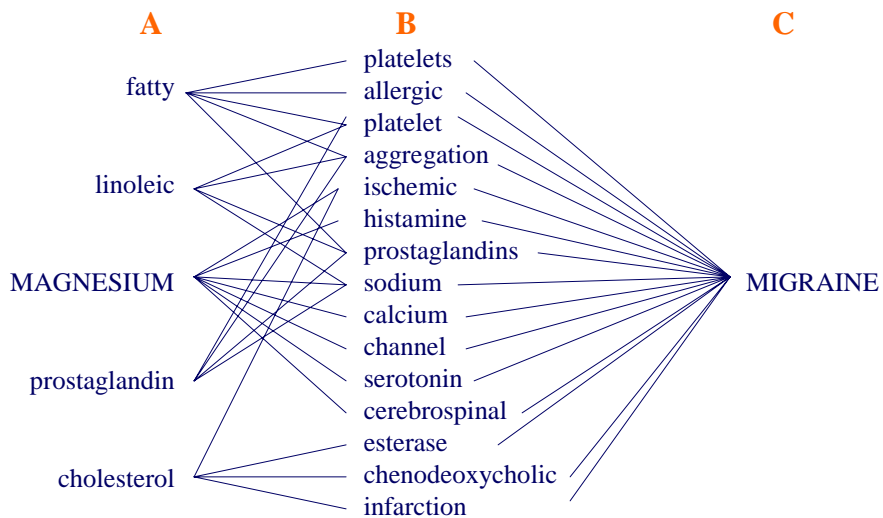
Task: Finding Complementary Literatures

- ARROWSMITH aids in identifying relationships that are implicit, but not explicitly described, in the literature
- <http://kiwi.uchicago.edu/>
- Swanson & Smalheiser, *Artificial Intelligence* 91, 1997

ARROWSMITH: Finding Complementary Literatures



ARROWSMITH Example: The Magnesium-Migraine Link



The ARROWSMITH Method

- given: query concept **C** (e.g. *migraine*)
- do:
 - run MEDLINE search on **C**
 - derive a set of words (**B**) from titles of returned articles; retain words
 - run MEDLINE search on each **B** word to assemble list of **A** words
 - rank **A-C** linkages by number of different intermediate **B** terms

Restricting the Search in ARROWSMITH

- prune **B** list by
 - using a predefined stop-list (“clinical”, “comparative”, “drugs”,...)
 - having a human expert filter terms
- prune **A** list using *category restrictions* (e.g. dietary factors, toxins, etc.)
- prune **C-B**, **B-A** linkages by requiring:
$$\Pr(B | C) > \Pr(B)$$
$$\Pr(A | B) > \Pr(A)$$

Given a document with word **C**, do we see **B** more often than we'd expect by chance?

ARROWSMITH Case Studies

- *indomethacin and Alzheimer's disease*
- *estrogen and Alzheimer's disease*
- *phospholipases and sleep*
- etc.

Task: Named Entity Recognition

- for many text analysis tasks it is useful to recognize that certain words/phrases describe a particular type of entity
 - genes
 - proteins
 - RNAs, cell types, strains, chromosomal locations, etc.
- Fukuda et al., *Pacific Symposium on Biocomputing*, 1998
 - recognizing protein names using *morphological*, *lexical*, and *syntactic* information

Named Entity Recognition

- relevant sources of information
 - *morphological*: how words are constructed from “parts”
 - *lexical*: specific words and word classes
 - *syntactic*: how words are composed into grammatical units

Recognizing Protein Names

- morphological 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...

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

The focal adhesion kinase (FAK) is...

Recognizing Protein Names: Morphological Analysis

- 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

- 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

- extend term to include a succeeding upper-case letter or a Greek-letter word

p85 alpha → p85 alpha

Task: Relation Extraction

- given: specific relations of interest
- do: extract instances of the relations from text sources
- Craven & Kumlien, *ISMB 99*; Ray and Craven, *IJCAI 2001*; Skounakis, Ray & Craven, *in preparation*

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 pleiotropic 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

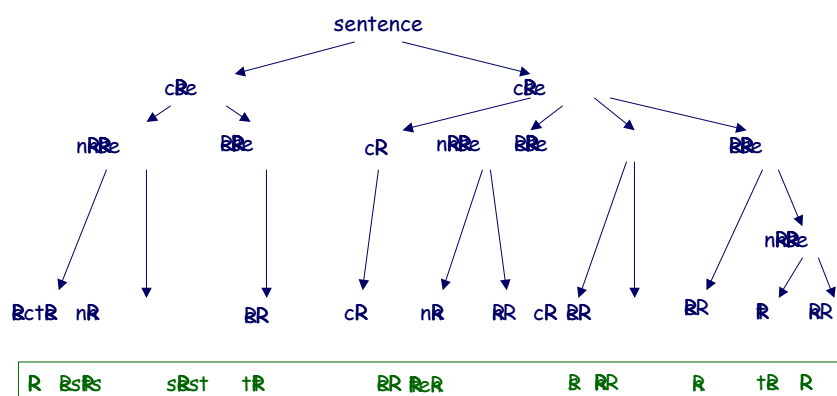
- providing structured summaries for queries
 - What is known about protein X (subcellular & tissue localization, associations with diseases, interactions with drugs, ...)?
- assisting in the construction and updating of databases
- assisting scientific discovery by detecting previously unknown relationships, annotating experimental data

Hidden Markov Models for Information Extraction

- there are efficient algorithms for doing the following with HMMs:
 - determining the likelihood of a sentence given a model
 - determining the most likely path through a model for a sentence
 - setting the parameters of the model to maximize the likelihood of a set of sentences

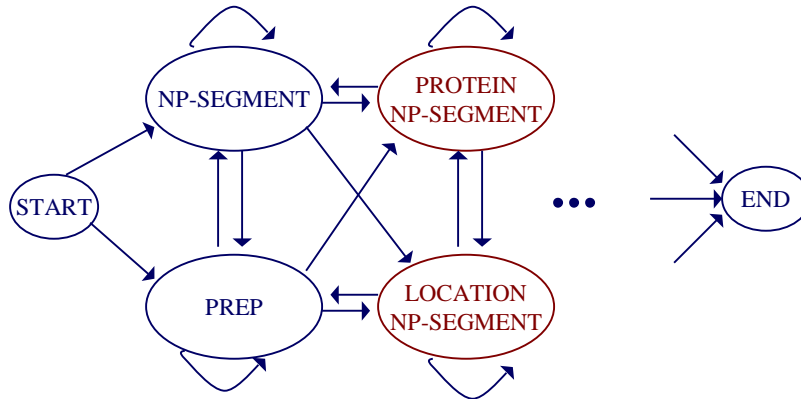
Representing Sentences

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



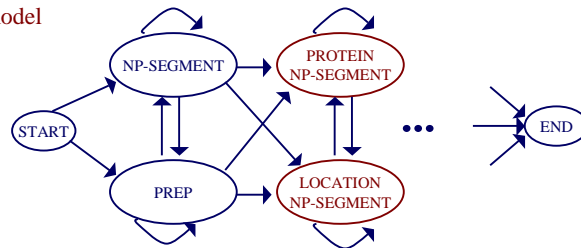
Multiple Resolution HMMs (Part 1)

- [Ray & Craven, *IJCAI 01*; Skounakis et al, *in preparation*]
- states have types, emit *phrases*
- some states have labels (**PROTEIN**, **LOCATION**)
- our models have ≈ 25 states at this level

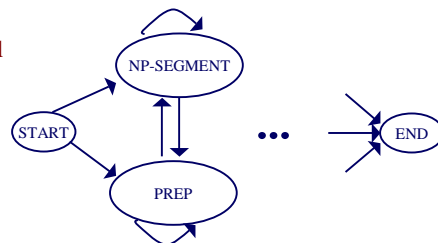


Multiple Resolution HMMs (Part 2)

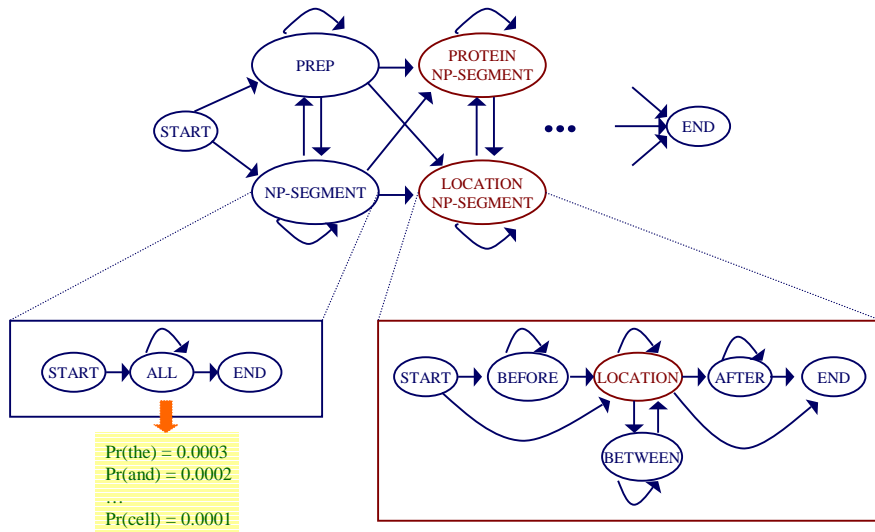
positive model



null model

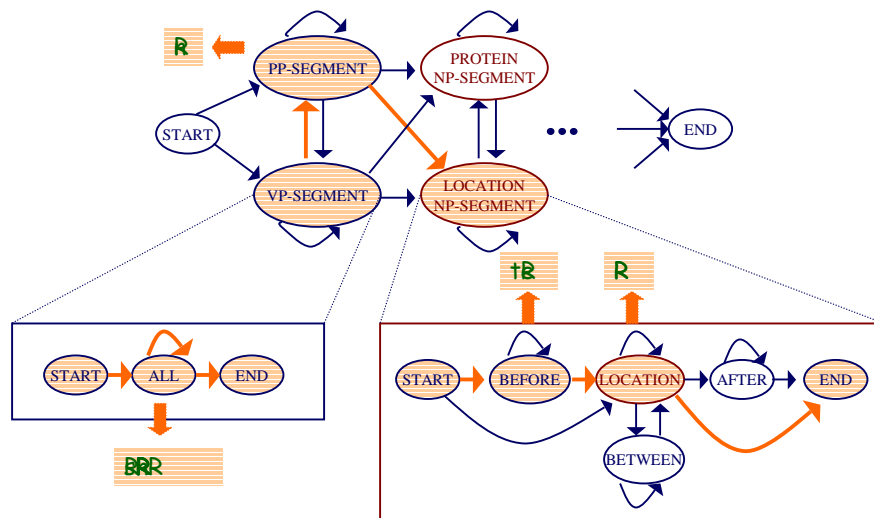


Multiple Resolution HMMs (Part 3)



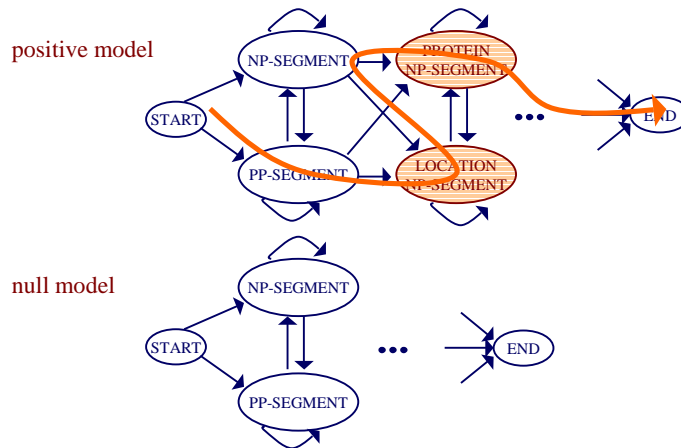
Multiple Resolution HMMs

- consider emitting: **RRRRR**



Extraction with our HMMs

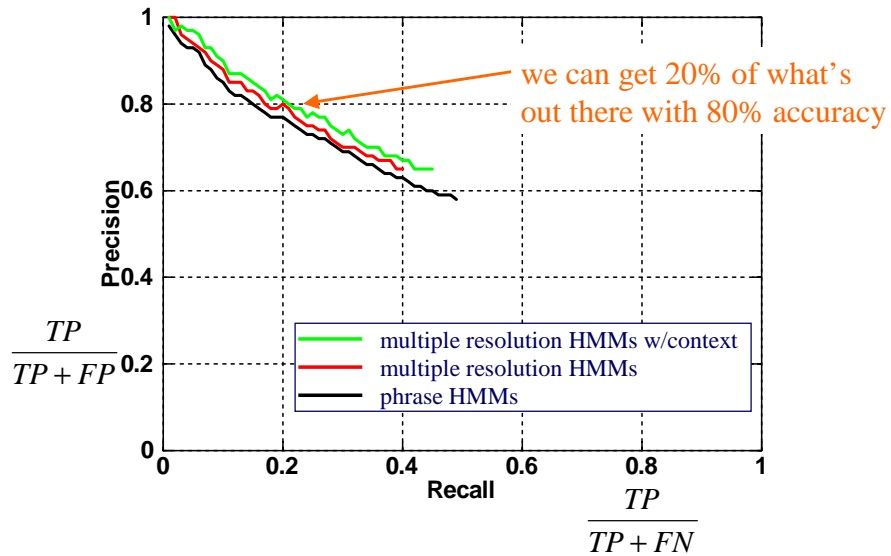
- extract a relation instance if
 - sentence is more probable under positive model
 - Viterbi path goes through special extraction states



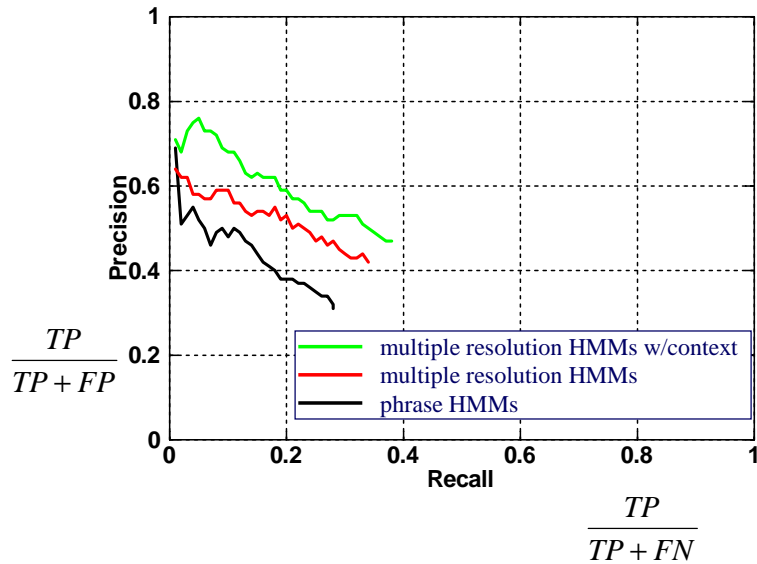
Data Sets

-)
 - YPD database
 - 769 positive, 6193 negative sentences
 - 939 tuples (402 distinct)
-)))
 - OMIM database
 - 829 positive, 11685 negative sentences
 - 852 tuples (143 distinct)
-)))
 - MIPS database
 - 5446 positive, 41377 negative
 - 8088 (819 distinct)

Extraction Accuracy (MIPS)



Extraction Accuracy (YPD)



Extraction Accuracy (OMIM)

