

# Text Mining for Biomedicine

Sophia Ananiadou

[Sophia.Ananiadou@manchester.ac.uk](mailto:Sophia.Ananiadou@manchester.ac.uk)

Yoshimasa Tsuruoka

[tsuruoka@is.s.u-tokyo.ac.jp](mailto:tsuruoka@is.s.u-tokyo.ac.jp)

# Outline

- Challenges of text mining in biomedicine
- Resources for text mining in biomedicine
- Terminology processing
- Information Extraction
- Levels of linguistic analysis

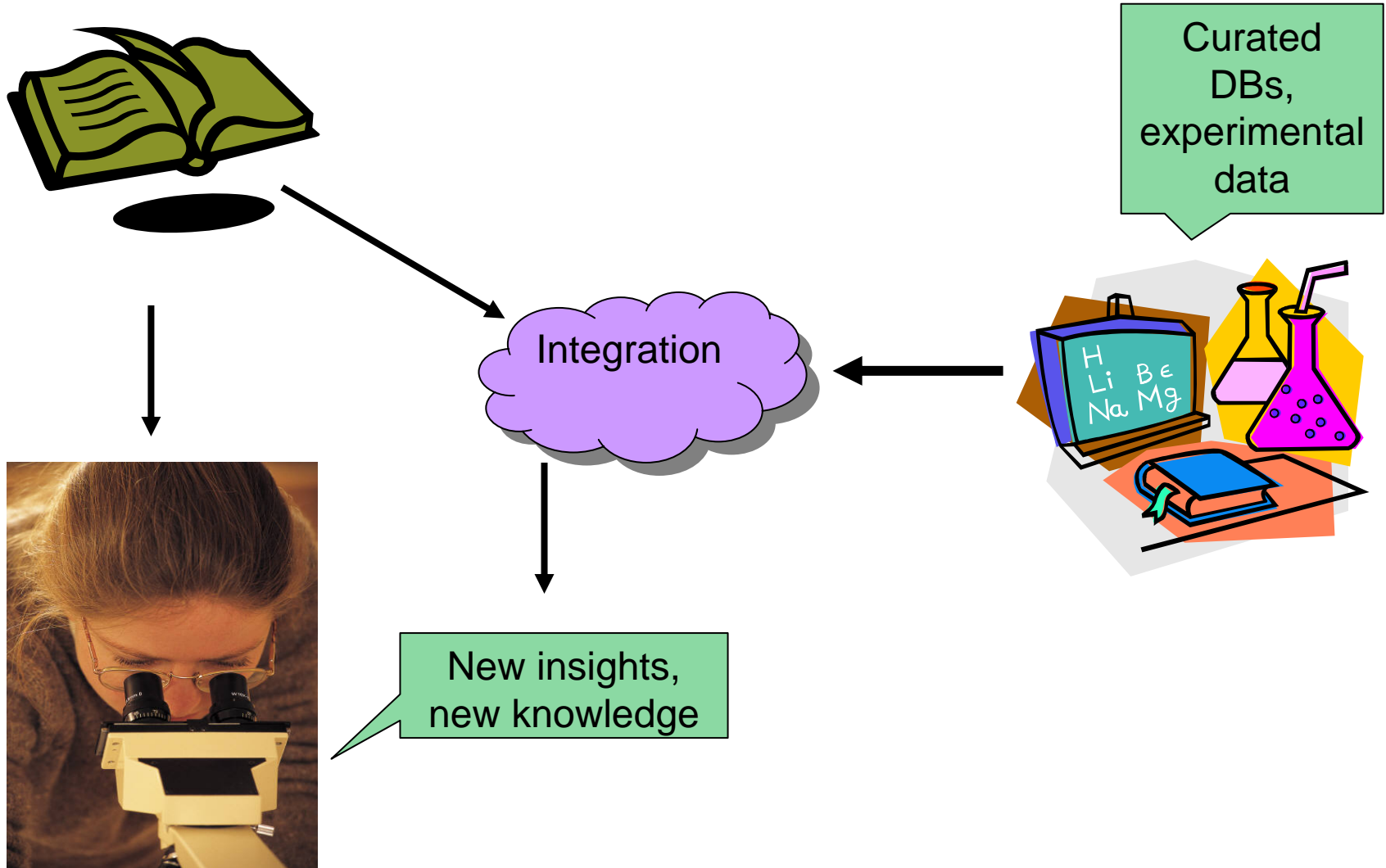
# Challenges of TM in biomedicine

- Why biomedicine?
  - Consider just MEDLINE: 16,000,000 references, 40,000 added per month
  - Dynamic nature of the domain: new terms (genes, proteins, chemical compounds, drugs) constantly created
  - Impossible to manage such an information overload

# Information deluge

- Bio-databases, controlled vocabularies and bio-ontologies encode only small fraction of information
- **Linking** text to databases and ontologies
  - Curators struggling to process scientific literature
  - Discovery of facts and events crucial for gaining insights in biosciences: need for text mining

# Gaining New Insights



# Aims

- **Text mining**: discover & extract **unstructured** knowledge hidden in text
  - Hearst (1999)
- Text mining aids to **construct hypotheses** from associations derived from text
  - protein-protein interactions
  - associations of genes – phenotypes
  - functional relationships among genes...etc

Swanson, D.

# Text mining steps

- **Information Retrieval** yields all relevant texts
  - Gathers, selects, filters documents that may prove useful
  - Finds what is known
- **Information Extraction** extracts facts & events of interest to user
  - Finds relevant concepts, facts about concepts
  - Finds only what we are looking for
- **Data Mining** discovers unsuspected associations
  - Combines & links facts and events
  - Discovers *new* knowledge, finds new associations

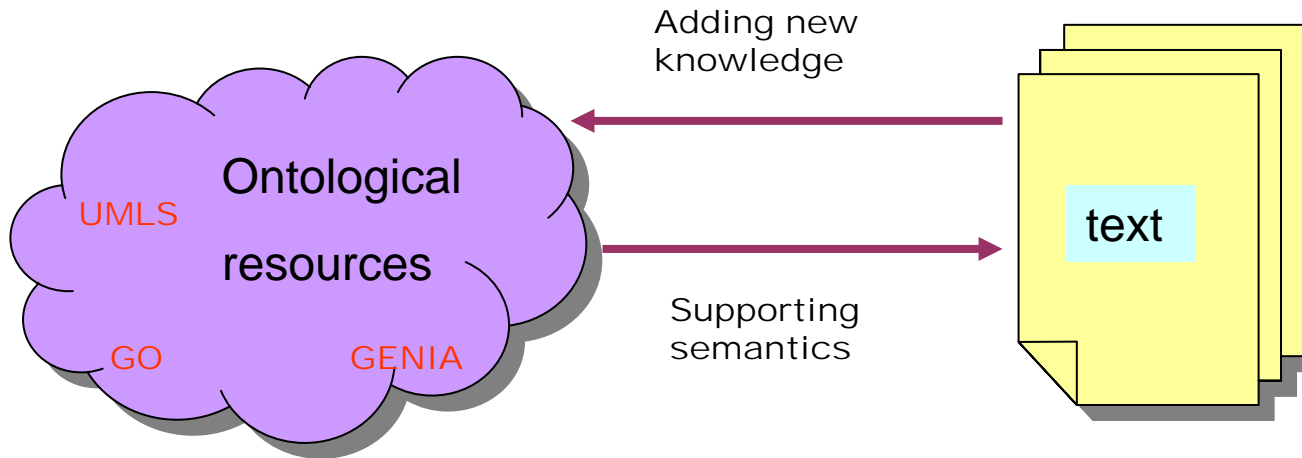
# Challenge: the resource bottleneck

- Lack of large-scale, richly annotated **corpora**
  - Support training of ML algorithms
  - Development of computational grammars
  - Evaluation of text mining components
- Lack of **knowledge resources**: lexica, terminologies, ontologies.

# What about existing resources?

- Ontologies important for knowledge discovery
  - They form the link between terms in texts and biological databases
  - Can be used to add **meaning**, semantic annotation of texts

# Link between text and ontologies



# Resources for Bio-Text Mining

- Lexical / terminological resources
    - SPECIALIST lexicon, Metathesaurus (UMLS)
    - Lists of terms / lexical entries (hierarchical relations)
  - Ontological resources
    - Metathesaurus, Semantic Network, GO, SNOMED CT, etc
    - Encode relations among entities
- ➡ Bodenreider, O. “Lexical, Terminological, and Ontological Resources for Biological Text Mining”, Chapter 3, Text Mining for Biology and Biomedicine, pp.43-66

# SPECIALIST lexicon

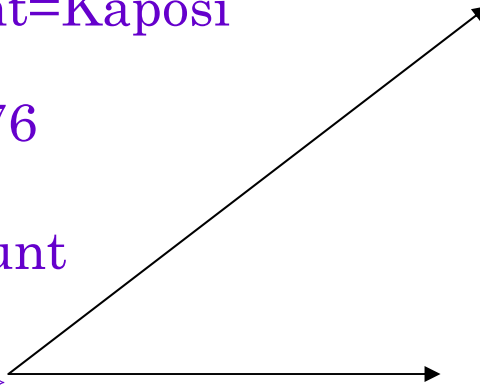
## – UMLS specialist lexicon

<http://SPECIALIST.nlm.nih.gov>

- Each lexical entry contains morphological (e.g. *cauterize, cauterizes, cauterized, cauterizing*), syntactic (e.g. complementation patterns for verbs, nouns, adjectives), orthographic information (e.g. *esophagus – oesophagus*)
- General language lexicon with many biomedical terms (over 180,000 records)
- Lexical programs include variation (spelling), base form, inflection, acronyms

# Lexicon record

{base=Kaposi's sarcoma  
spelling\_variant=Kaposi  
sarcoma  
entry=E0003576  
cat=noun  
variants=uncount  
variants=reg  
variants=glreg}



**Kaposi's sarcoma**

**Kaposi's sarcomas**

**Kaposi's sarcomata**

**Kaposi sarcoma**

**Kaposi sarcomas**

**Kaposi sarcomata**

***The SPECIALIST Lexicon and Lexical Tools***

Allen C. Browne, Guy Divita, and Chris Lu PhD

2002 NLM Associates Presentation, 12/03/2002, Bethesda, MD

# Normalisation (lexical tools)

Hodgkin Disease

HODGKIN DISEASE

Hodgkin's Disease

Hodgkin's disease

Disease, Hodgkin ...

normalise



disease hodgkin

# Steps of Norm

Remove genitive

Hodgkin's Diseases

Replace punctuation with spaces

Hodgkin Diseases

Remove stop words

Hodgkin Diseases

Lowercase

hodgkin diseases

Uninflect each word

hodgkin disease

Word order sort

disease hodgkin

**Lexical tools of the UMLS**

<http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html>

# The Gene Ontology (GO)

- Controlled vocabulary for the annotation of gene products
  - <http://www.geneontology.org/>
  - 19,468 terms. 95.3% with definitions (as of 4/2/2006)
    - 10391 biological\_process
    - 1681 cellular\_component
    - 7396 molecular\_function

# Gene Ontology

- GOA database (<http://www.ebi.ac.uk/GOA/>) assigns gene products to the Gene Ontology
- GO terms follow certain conventions of creation, have synonyms such as:
  - *ornithine cycle* is an exact synonym of *urea cycle*
  - *cell division* is a broad synonym of *cytokinesis*
  - *cytochrome bc1 complex* is a related synonym of *ubiquinol-cytochrome-c reductase activity*

## GO terms, definitions and ontologies in OBO

id: GO:0000002

name: mitochondrial genome maintenance

namespace: biological\_process

def: "The maintenance of the structure and integrity of the mitochondrial genome."

[GOC:ai]

is\_a: GO:0007005 ! mitochondrion organization and biogenesis

[http://www.geneontology.org/ontology/gene\\_ontology.obo](http://www.geneontology.org/ontology/gene_ontology.obo)

**id:** GO:0000024

**name:** maltose biosynthesis

**namespace:** biological\_process

**def:** "The formation from simpler components of the disaccharide maltose (4-O-alpha-D-glucopyranosyl-D-glucopyranose)." [GOC:jl, ISBN:0198506732]

**subset:** gosubset\_prok

**exact\_synonym:** "malt sugar biosynthesis" []

**exact\_synonym:** "maltose anabolism" []

**exact\_synonym:** "maltose formation" []

**exact\_synonym:** "maltose synthesis" []

**is\_a:** GO:0000023 ! maltose metabolism

**is\_a:** GO:0046351 ! disaccharide biosynthesis

Names, synonyms,  
relations used for TM

# Metathesaurus

- organised by concept
  - 5M names, 1M concepts, 16M relations
- built from 134 electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms
- "source vocabularies"
- common representation

# Are existing knowledge resources sufficient for TM?

1. Limited lexical & terminological coverage of biological sub-domains
2. Resources focused on human specialists  
GO, UMLS, UniProt ontology concept names frequently confused with terms

## Occurrences of GO 'terms' in text

53,000 abstracts (baker's yeast) contained only 8,000 occurrences of 739 distinct GO terms

# Naming conventions

## 3. Update and curation of resources

- FlyBase gene name coverage 31% (abstracts) to 84% (full texts)

## 4. Naming conventions and representation in heterogeneous resources

- Term formation guidelines from formal bodies e.g. HUGO, IPI not uniformly used
- Problems with integration of resources

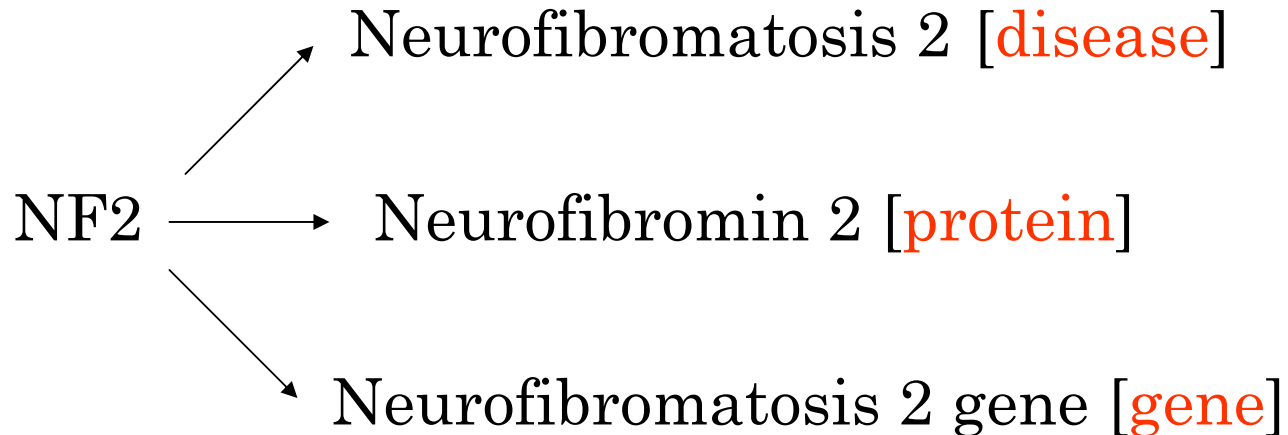
*dystrophin* used for 18 gene products

“*Dystrophin (muscular dystrophy, Duchenne and Becker types), included DXS143, DXS164, DXS206, ...*” HUGO

# Term variation

5. Terminological variation and complexity of names
  - High correlation between degree of term variation and dynamic nature of biomedicine
  - Variation occurs in controlled vocabularies and texts but discrepancy between the two
  - Exact match methods fail to associate term occurrences in texts with databases
  - Mouse gene names, variations accounted for 79% missing gene occurrences (Tuason, 2004)

# Term ambiguity



O. Bodenreider, MIE 2005 tutorial

<http://www.nactem.ac.uk/>

# Term ambiguity

- Gene terms may be also common English words
  - *BAD* human gene encoding BCL-2 family of proteins (*bad news, bad prediction*)
- Gene names are often used to denote gene products (proteins)
  - *suppressor of sable* is used ambiguously to refer to either genes and proteins
- Existing resources lack information that can support term disambiguation
- Difficult to establish equivalences between termforms and concepts

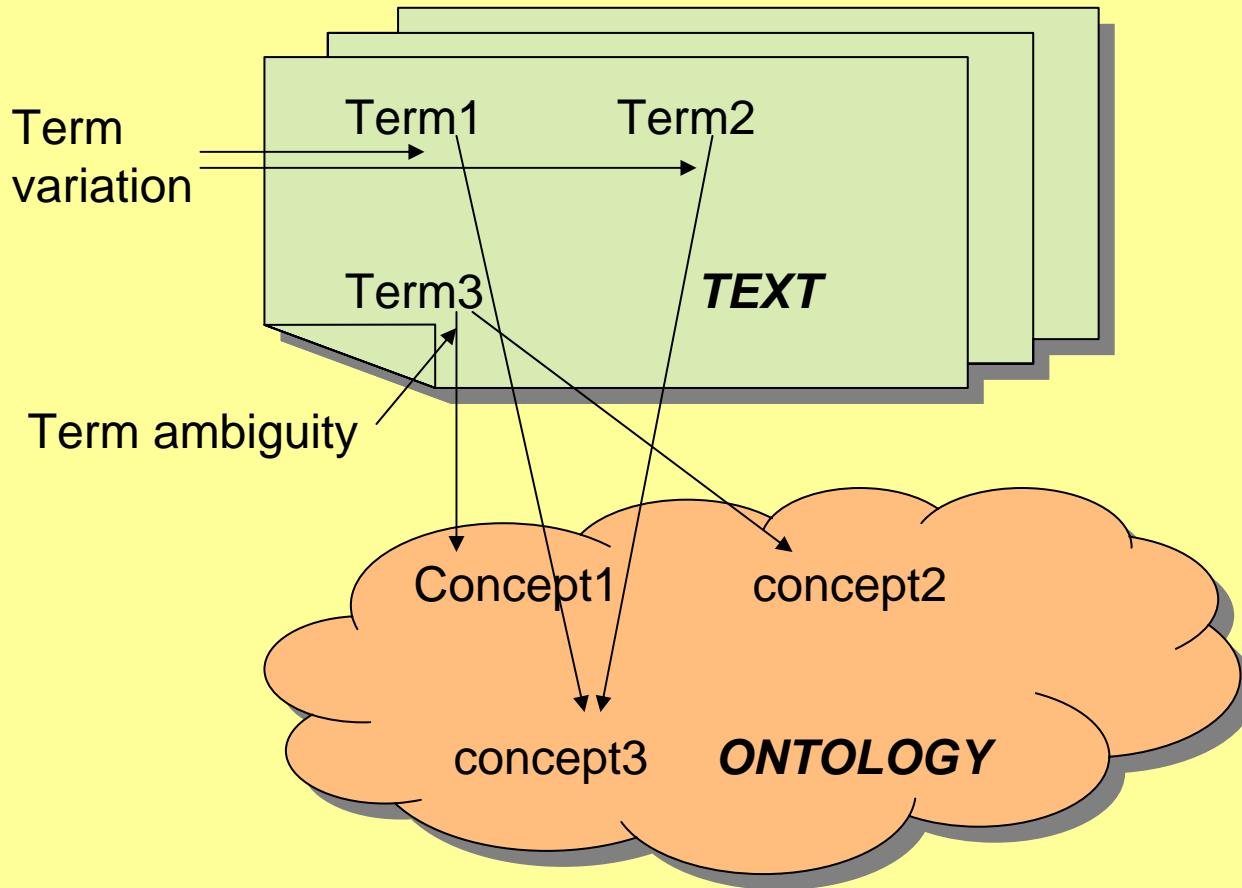
# Homologues

- **Cycline-dependent kinase inhibitor** first introduced to represent a protein family **p27**
  - But it is used interchangeably with **p27** or **p27kip1**, as the name of the individual protein and not as the name of the protein family (Morgan 2003).
- **NFKB2** denotes the name of a family of 2 individual proteins with separate IDs in Swiss-Prot.
  - These proteins are homologues belonging to different species, homo sapiens & chicken.

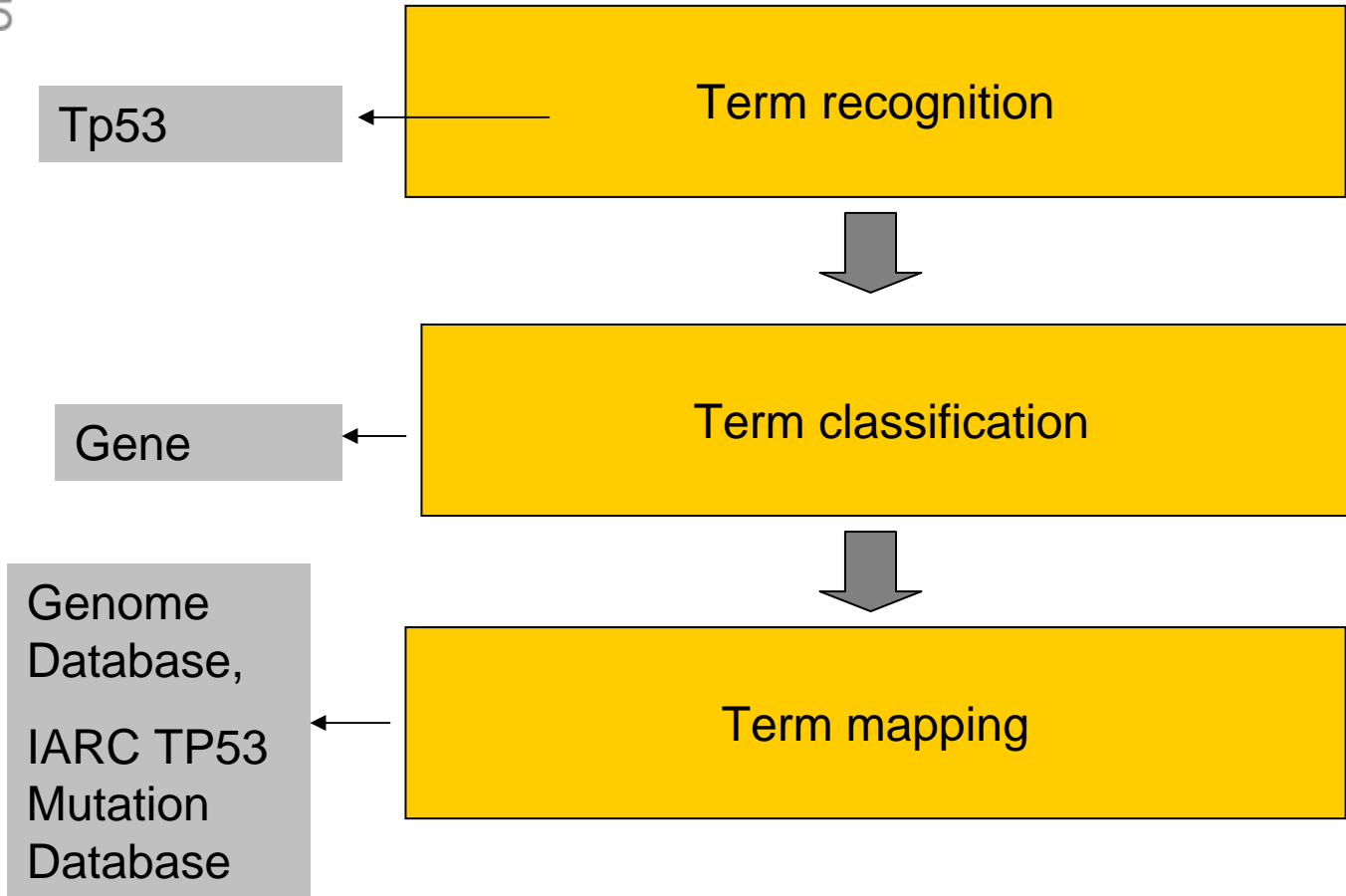
# Terms

- **Term**: linguistic realisation of specialised concepts, e.g. genes, proteins, diseases
- **Terminology**: collection of terms structured (hierarchy) denoting relationships among concepts, part-whole, is-a, specific, generic, etc.
- Terms link text and ontologies
- Mapping is not trivial (main challenge)

# Term variation and ambiguity



# Term mining steps



# Term recognition techniques

- **ATR** extracts terms (variants) from a collection of document
- Distinguishes terms vs non-terms
- In **NER** the steps of **recognition and classification** are merged, a classified terminological instance is a named entity
- The tasks of ATR and NER share techniques but their ultimate goals are different
  - ATR for resource building, lexica & ontologies
  - NER first step of IE, text mining

# Overview papers

1. S. Ananiadou & G. Nenadic (2006) Automatic Terminology Management in Biomedicine, Text Mining for Biology and Biomedicine, pp. 67- 97.
2. M. Krauthammer & G. Nenadic (2004) Term identification in the biomedical literature, JBI 37 (2004) 512-526
3. J.C. Park & J. Kim (2006) Named Entity Recognition, Text Mining for Biology and Biomedicine, pp. 121-142

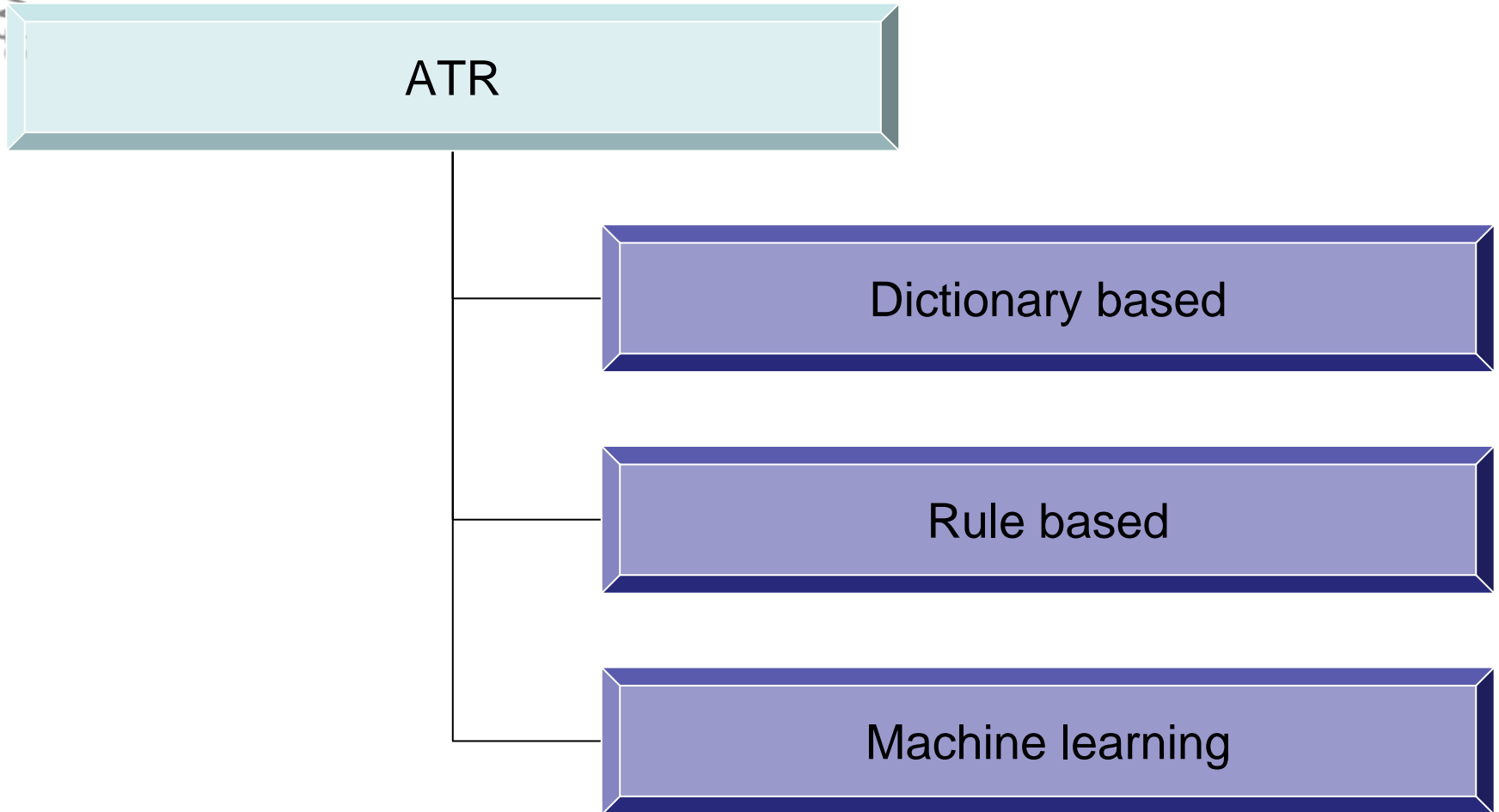
## Detailed bibliography in Bio-Text Mining

1. BLIMP <http://blimp.cs.queensu.ca/>
2. <http://www.ccs.neu.edu/home/futrelle/bionlp/>

## Book on BioText Mining

1. S. Ananiadou & J. McNaught (eds) (2006) Text Mining for Biology and Biomedicine, Artech House.

# Main ATR approaches



# Dictionary NER (1)

- Use terminological resources to locate term occurrences in text
  - NCBI <http://www.ncbi.nlm.nih.gov/>
  - EBI <http://www.ebi.ac.uk/>
  - neologisms, variations, ambiguity problematic for simple dictionary look-up
  - Ambiguous words e.g. *an*, *for*, *can* ...
  - spelling variants, punctuation, word order variations
    - *estrogen oestrogen*
    - *NF kappa B / NF kB*

# Dictionary NER (2)

- Hirschman (2002) used FlyBase for gene name recognition, results disappointing due to **homonymy, spelling variations**
  - Precision, 7% abstracts, 2% full papers
  - Recall, 31% -- 84%
- Tuason (2004) reports term variation as main problem of mismatch
  - *bmp\_4*      *bmp4*
  - *syt4*      *syt iv*
  - *integrin alpha 4*      *alpha4 integrin*

# Dictionary NER (3)

- Krauthammer (2000) use string comparison for gene & protein name recognition
- Uses nucleotide combination {A, C, G, T} to convert text
- Applies BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ools), a set of sequence comparison algorithms, to text to identify similarities of gene and protein names
- Recall 78.8% and precision 71%

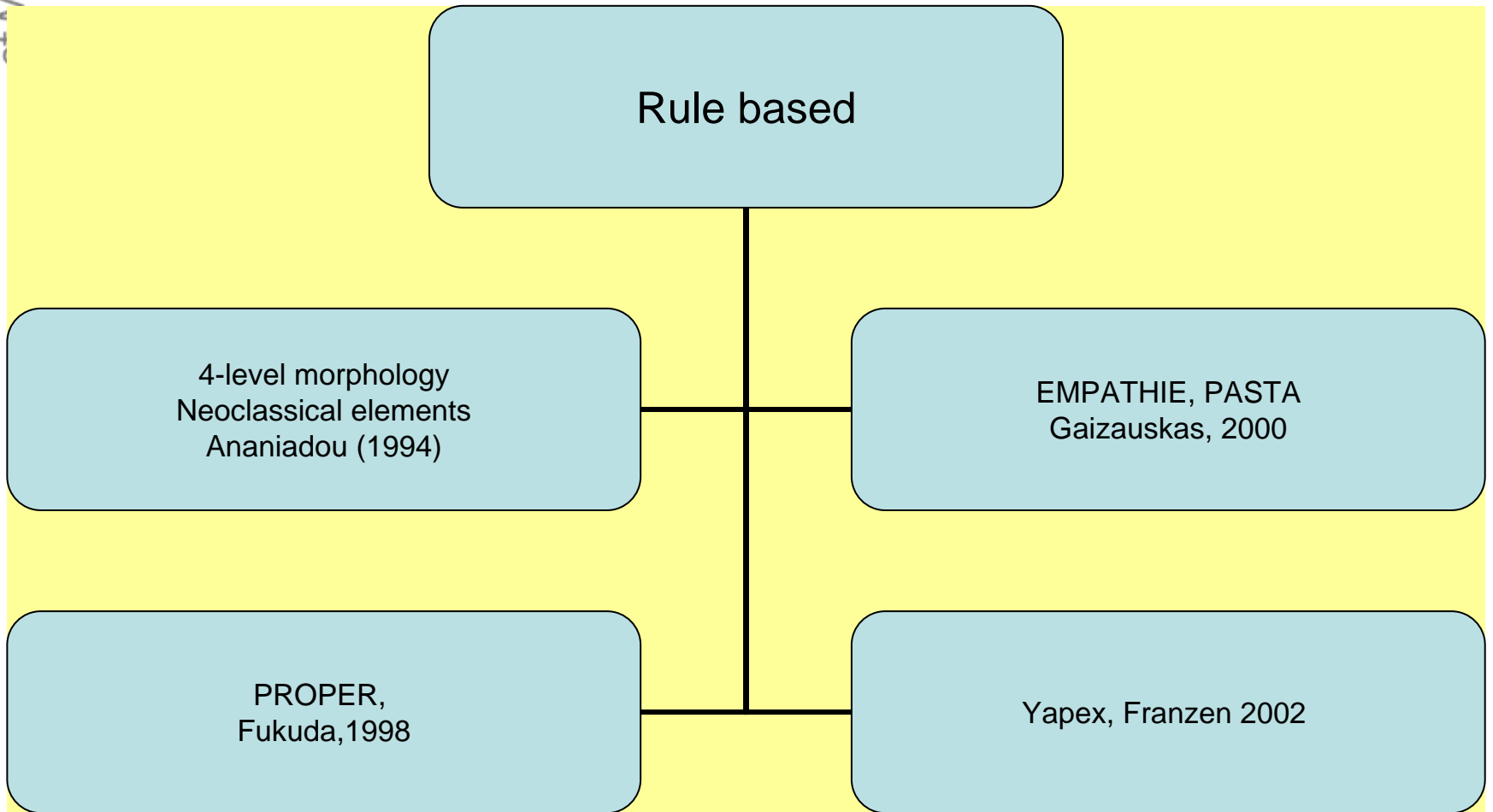
# Dictionary NER (4)

- Tsuruoka & Tsujii (2003) suggest a probabilistic generator of spelling variants, edit distance operations (delete, substitute, insert)
  - Terms with  $ED \leq 1$  considered spelling variants
  - Used a dictionary of protein terms
- Support query expansion
- Augment dictionaries with variation

# Dictionary NER (5)

- TERMINO (Harkema 2004) stores biomedical terminological information, collected from various resources (e.g. UMLS, GOA, etc)
- attempts to establish and maintain links between resources of various types (ontology concept names, terms, controlled vocabularies, nomenclatures, classification descriptors)
- uses an external analyser to handle regular morphology
  - does not account for special morphological behaviour of terms

# Rule NER (2)



# Rule based (1)

- Use orthographic, morpho-syntactic features of terms
  - Rules that make use of internal term formation patterns (tagging, morphological analysers) e.g. affixes, combining forms
  - Do not take into account contextual features
  - Dictionaries of constituents e.g. affixes, neoclassical forms included
- Portability to different domains?

# Rule based (2)

- Ananiadou, S. (1994) recognised single-word terms based on morphological analysis of term formation patterns (internal term make up)
- based on analysis of neoclassical and hybrid elements
  - ‘alphafetoprotein’ ‘immunoosmoelectrophoresis’
  - ‘radioimmunoassay’
- some elements are used for creating terms
  - term → word + term\_suffix
  - term → term + word\_suffix
- neoclassical combining forms (electro- adeno-),
- prefixes (auto-, hypo-)
- suffixes ( -osis, -itis)

# Rule-based (3)

- Fukuda (1998) used lexical, orthographic features for protein name recognition e.g. upper case character, numerals etc.
- PROPER: **core** and **feature** elements
  - Core: meaning bearing elements
  - Feature: function elements

core → **SAP kinase** ← feature

Core elements extended to feature based on concatenation rules (based on POS tags)

# Rule-based (4)

- Gaizauskas (2000) CFG for protein name recognition (PASTA, EMPATHIE)
- Based on morphological and lexical characteristics of terms
  - biochemical suffixes (-ase enzyme name)
  - dictionary look-up (protein names, chemical compounds, etc)
  - deduction of term grammar rules from Protein Data Bank

Protein -> protein\_modifier, protein\_head, numeral

# Rule-based (5)

- Inspired by PROPER, Yapex uses Swiss-Prot to add core term elements

<http://www.sics.se/humle/projects/prothalt/yapex.cgi>

- Hou (2003) used Yapex with context information (collocations) appearing with protein names
- Rule based approaches construct rule and patterns manually or automatically
- Difficult to tune to different domains

# Machine learning systems

- Learn features from training data for term recognition and classification
- Most ML systems combine recognition and classification

## Challenges

- Feature selection and optimisation
- Availability of training data
- detection of term boundaries

# ML (1)

- Collier (2000) used HMM, orthographic features for term recognition
  - HMM looks for most likely sequence of classes corresponding to a word sequence e.g. **interleukin-2 protein/DNA**
  - To find similarities between known words (training set) and unknown words, use character features

## Feature

DigitNumber

GreekLetter

TwoCaps

## Examples

[2]protein[3]DNA

[alpha]protein

[RelB]protein[TAR]RNA

# ML (2)

- Use of GENIA resources as training data
  - Results depend on training data
- Morgan (2004) used FlyBase to construct automatically training corpus
  - Pattern matching for gene name recognition, noisy corpus annotated
  - HMM was trained on that corpus for gene name recognition

# Support Vector Machines (1)

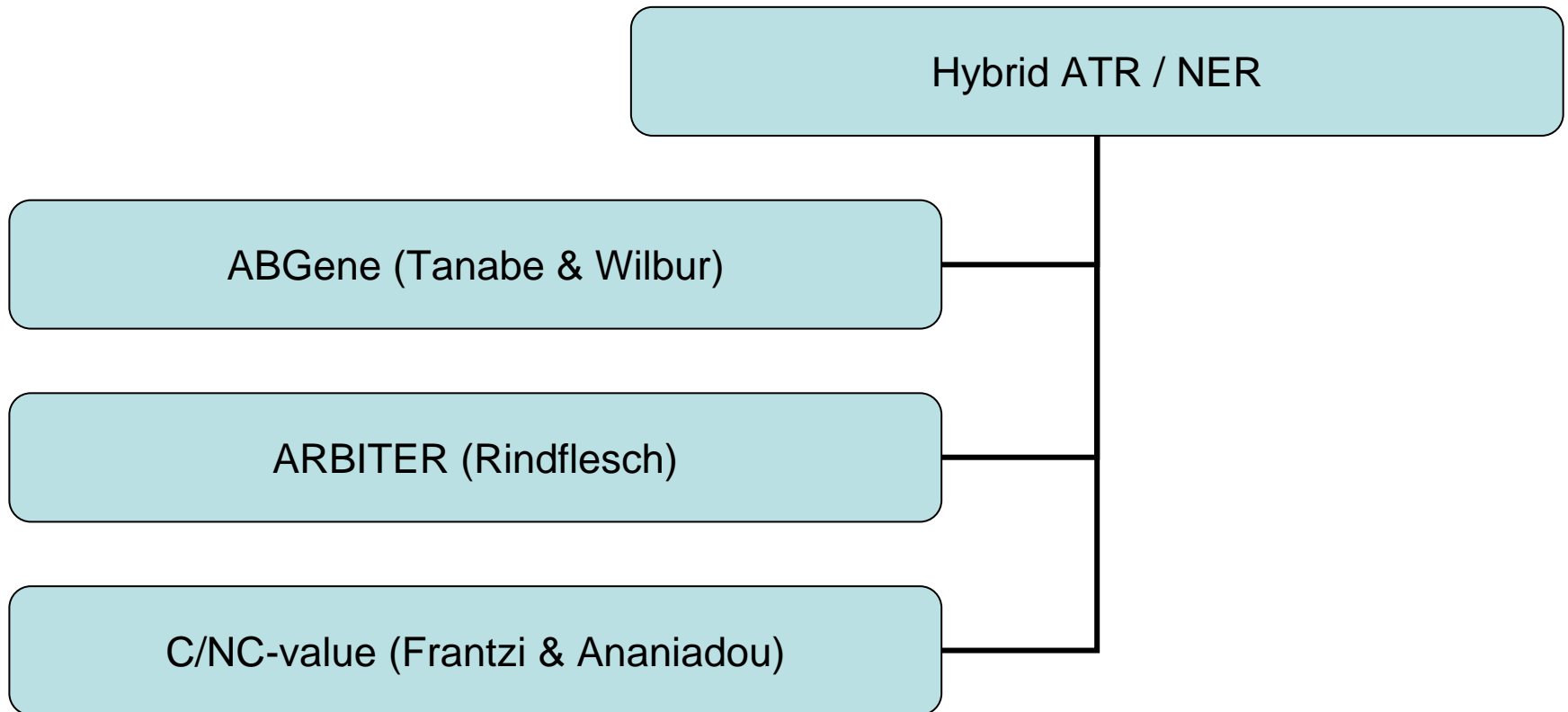
- Kazama trained multi-class SVMs on Genia corpus
- Corpus annotated with B-I-O tags
  - B tags denote words at beginning of term
  - I tags inside term
  - O tags outside term
  - **B-protein-tag** : word in the beginning of a protein name

# SVMs for NER (2)

- Yamamoto used a combination of features for protein name recognition:
  - Morphological, lexical, boundary, syntactic (head noun), *domain specific* (if term exists in biomedical database).
- Lee use different features for **recognition** and **classification**.
  - orthographic, prefix, suffix
  - **Contextual information**

# Hybrid approaches

- Combine rules, statistics, resources



# Hybrid (1)

- ABGene: protein and gene name tagger
  - Combines ML, transformation rules, dictionaries with statistics
  - Protein tagger trained on MEDLINE abstracts by adapting Brill's tagger
  - Transformation rules for recognition of gene, protein names
  - Used GO, LocusLink list of genes, proteins for false negative tags

# Hybrid (2)

- ARBITER (Access and Retrieve Binding Terms) uses
  - UMLS Metathesaurus and GenBank to map NPs (binding terms)
  - morphological features
  - lexical information (head noun)
- EDGAR recognises gene, cell, drug names using co-occurrences of **cell, clone, expression**

# Hybrid (3)

- C/NC value Frantzi & Ananiadou, 1998
- C-value
  - Linguistic filters
  - total frequency of occurrence of string in corpus
  - frequency of string as part of longer candidate terms (nested terms)
  - number of these longer candidate terms
  - length of string (in number of words)
- Output: automatically ranked terms

# C-value

- **C-value measure** extracts multi-word, nested terms

[adenoid [cystic [basal [cell carcinoma]]]]

cystic basal cell carcinoma

ulcerated basal cell carcinoma

recurrent basal cell carcinoma

basal cell carcinoma

# Term variation

- variation recognition as *part of ATR*
- recognise term forms and link them into equivalence classes
- important if ATR is based on statistics (e.g. frequency of occurrence)
  - corpus-based measures are distributed across different variants
  - conflation of various surface representations of a given term should improve ATR

# Simple variation

- orthographic
  - hyphens, slashes (*amino acid* and *amino-acid*)
  - lower/upper cases (*NF-KB* and *NF-kb*)
  - spelling variations (*tumour* and *tumor*)
  - transliterations (*oestrogen* and *estrogen*)
- morphological
  - inflectional phenomena (plural, possessives)
- lexical
  - genuine synonyms (*carcinoma* and *cancer*)

# Complex variation

- Structural
  - Possessive usage of nouns using prepositions (*clones of human* and *human clones*)
  - Prepositional variants (*cell in blood*, *cell from blood*)
  - Term coordinations (*adrenal glands and gonads*)

# Coordinated term variants

- Structure is ambiguous
  - Head coordination or term conjunction?

|                   |   |
|-------------------|---|
| example           | <i>adrenal glands <u>and</u> gonads</i>     |
| head coordination | <i>[adrenal [glands <u>and</u> gonads]]</i> |
| term conjunction  | <i>[adrenal glands] <u>and</u> [gonads]</i> |

- Head or argument coordination?
  - $(N|A)^+ CC (N|A)^* N^+$ 
    - cell differentiation and proliferation
    - chicken and mouse receptors

# Nested terms

- Majority of terms are multi-word units
- Maximal vs nested term
  - [leukaemic [T [cell line]] Kit225]
- Recognising boundaries of multi-word terms important for NER
  - Spotting nested terms on their own in corpus not sufficient

# Nested terms

- Challenge here for ATR to identify the sub-strings which themselves are terms
  - Can help in resolving ambiguities in term variants
  - Inner term structure

*[leukaemic [T [cell line]]]*

*[leukaemic [[T and B] [cell lines]]]*

*⇒ leukaemic T cell line, leukaemic B cell line*

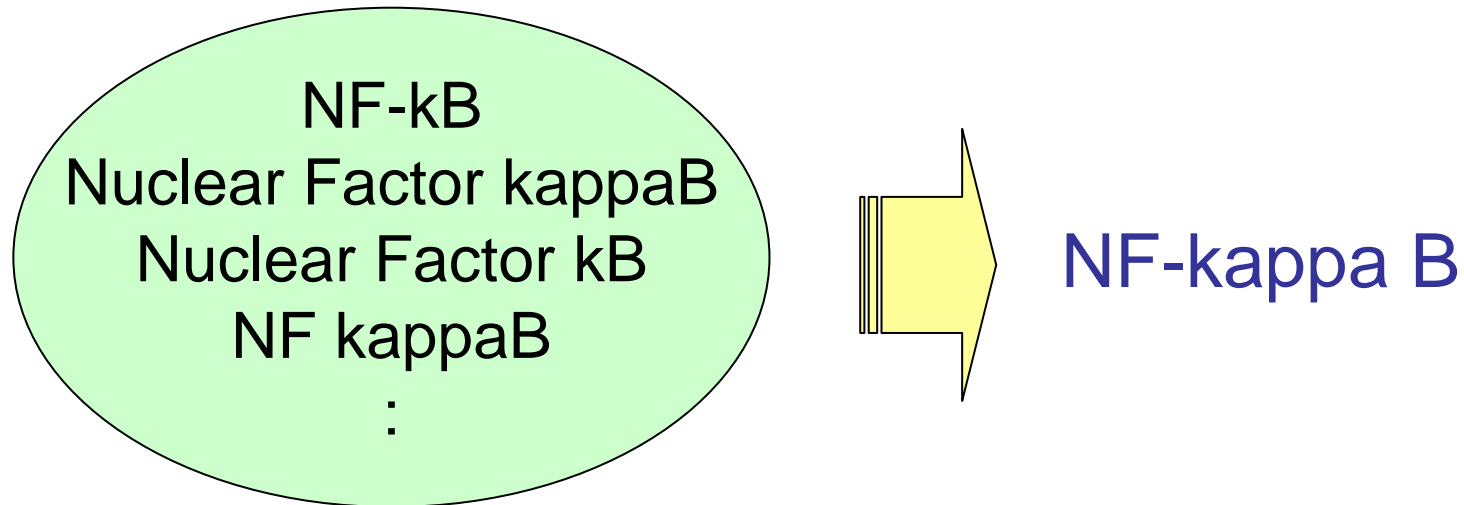
*⇒ leukaemic T cell line, B cell line*

# Acronyms

- Very productive type of term variation
- Acronym variation (synonymy)
  - NF kappa B/ NF kB / nuclear factor kappa B
- Acronym ambiguity (polysemy) even in controlled vocabularies



# Acronym variation



- Term variation is a big obstacle in knowledge integration. → Internal similarity of terms (edit-distance), spelling variation generator based on a probabilistic model, etc.

# Acronym recognition

- Swartz, A. & Hearst, M. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text, PSB 2003,8, 451-462
- Adar, E. (2004) SaRAD: a simple and robust abbreviation dictionary, Bioinformatics, 20(4) 527-533
- Chang, J.T. & Schutze, H. (2006) [Abbreviations in biomedical text, \*Text Mining for Biology and Biomedicine\*, pp.99-119, Artech](#)
- Nenadic, G., Spasic, I. & Ananiadou, S. (2002) Automatic acronym acquisition and term variation management, LREC, 2155-2162
- Tsuruoka, Y., Ananiadou, S. & Tsujii, J. (2005) A Machine learning approach to automatic acronym generation, ISMB, BioLink SIG, 25-31
- Pustejovsky, J. et al. (2001) Automatic extraction of acronym-meaning pairs from Medline databases, Medinfo, 10, 371-375.

# Recognition

- Extracting pairs of short and long forms
  - <acronym, long form>
  - Distinguishing acronyms from parenthetical expressions
  - Search for parentheses in text; single or more words; e.g. *Ab (antibody)*
  - Limit context around ( ); limit number of words according to number of letters in acronym

# Recognition

– Heuristics: match letters of acronym with letters of long form using rules, patterns

- letters from beginning of words
- combining forms

**c**arboxi**f**luorescein **d**iacetate (CFDA)

- Acronym normalisation to allow orthographic, structural and lexical variations (Nenadic et al)
- morphological information, positional info
- Penalise words in long form that do not match acronym
- Accidental matching

argininosuccitate synthet**a**se (AS)

The diagram shows the word 'argininosuccitate synthetase (AS)'. The letters 'a' and 's' in 'synthetase' are highlighted in red. Below the word, the letter 'A' has an arrow pointing to the red 'a', and the letter 'S' has an arrow pointing to the red 's'. This illustrates how the acronym 'AS' matches the letters 'a' and 's' in the long form word.

# Acronyms

- Alignment: find all matches between letters of acronyms and their long forms and calculate likelihood (Chang & Schütze)
    - Solves problem of acronyms containing letters not occurring in LF
    - LF includes words which do not have letters in acronyms, different word order
- <ADRB2, beta 2 adrenergic receptor>
- Choose best alignment based on features, e.g. position of letter etc.
  - Finding optimal weight for each feature challenge

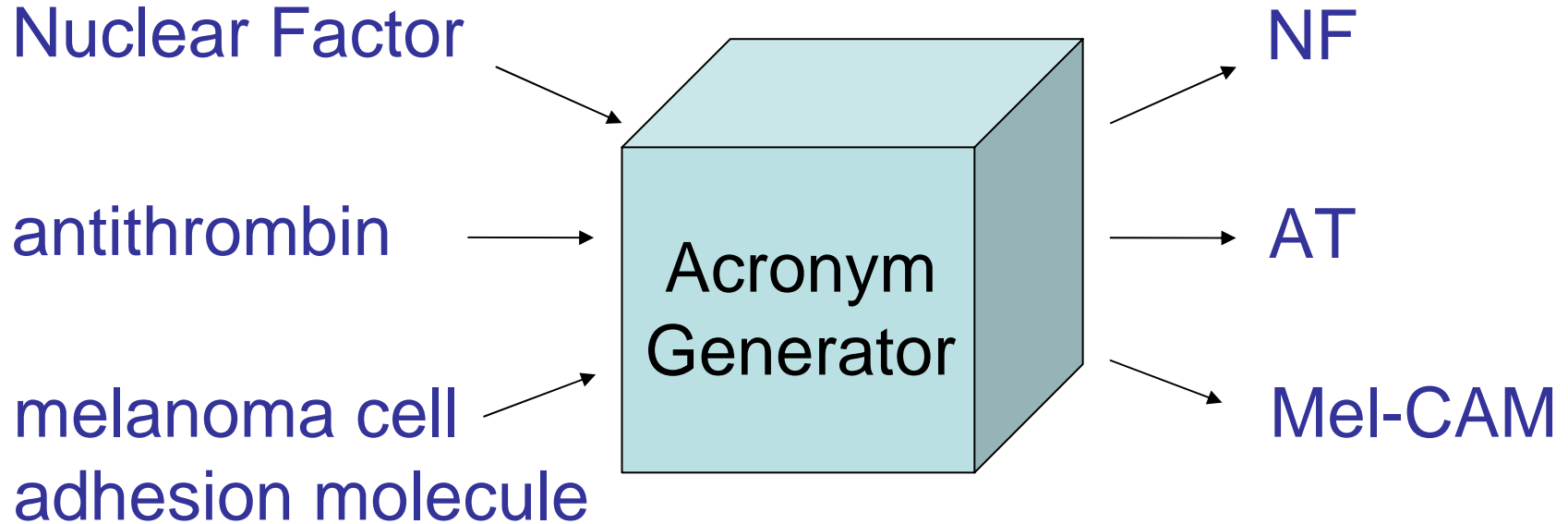
<http://abbreviation.stanford.edu/>

# Evaluating acronym recognition

- Acronyms in biomedicine are specific
  - General language recognisers perform poorly
    - Domain area covered by gold standard important; standards accurate, product of multiple experts
    - Medstract acronym gold standard

<http://www.medstract.org/gold-standards.html>

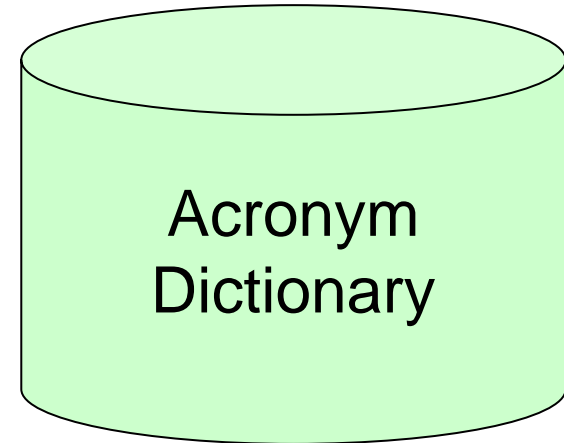
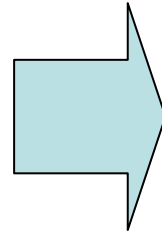
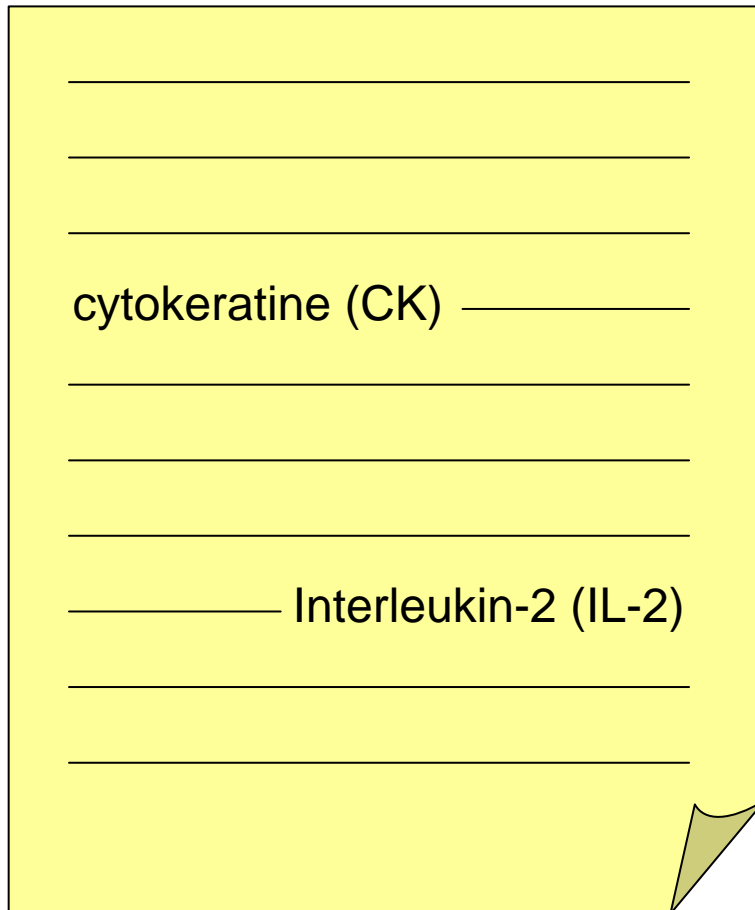
# Acronym Generation



- The system generates possible acronyms from a given expanded form.

# Dictionary-Building Approaches

## Running text



- Collect acronym-definition pairs from running text and construct a dictionary.

# Problems of Dictionary-Building Approaches

- Coverage
  - Limited available resources (corpora) and lack of generalization
  - Dynamic nature of terms
- Term variation in expanded forms
  - We need to address the problems of term variations in which acronyms are mixed with other variations such as spelling, lexical variations, etc.

# Acronym generator

- Machine learning-based
  - Acronym generation as sequence
  - Probabilistic modeling
- Advantages
  - Wide coverage can be achieved by generalization.
  - Similarities can be computed in a probabilistic form.
- Drawbacks
  - Needs training data

**MEMM** can capture features that reflect intuition of rule-based methods with statistical modeling

**Collection of weak cues**

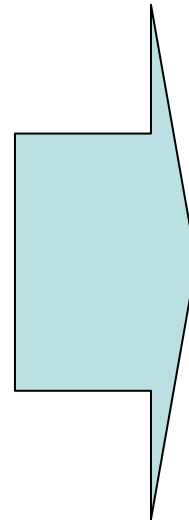
Tsuruoka, Y., Ananiadou, S. & Tsujii (2005) BioLink

# Acronym Generation as Sequence Tagging

cytokeratines

CKs

| Definition | Tag   |
|------------|-------|
| c          | UPPER |
| y          | SKIP  |
| t          | SKIP  |
| o          | SKIP  |
| k          | UPPER |
| e          | SKIP  |
| r          | SKIP  |
| a          | SKIP  |
| t          | SKIP  |
| i          | SKIP  |
| n          | SKIP  |
| e          | SKIP  |
| s          | LOWER |



| Acronym  |
|----------|
| <b>C</b> |
|          |
|          |
|          |
| <b>K</b> |
|          |
|          |
|          |
|          |
|          |
|          |
| <b>S</b> |

# Sequence Tagging with MEMM

## Maximum Entropy Modeling with Inequality Constraints (Kazama and Tsujii 2003, 2005)

- Smoothing effects  
Performance is better or comparable to that achieved with the use of Gaussian prior.
- Smaller model size -> quick decoding  
Ex. ) POS tagging
  - Gaussian prior: 12MB
  - Inequality constraints: 1.3MB

**MEMM** can integrate features that reflect intuition of rule-based methods with statistical modeling

maximum entropy classifier  
(model size = 60kB)

# Term Disambiguation

- Rindflesch used rules based on semantic types of neighbouring words when mapping terms to UMLS concepts;
  - Rule-based expensive and difficult to be comprehensive
- Use of supervised ML techniques but required manual annotation, expensive
- Liu et al. used a supervised approach to acquire semantically annotated corpora automatically based on UMLS
- Evaluation performed on ambiguous biomedical acronyms

# Disambiguation

- ML approaches (Naïve Bayesian, decision trees, etc) used to disambiguate gene, protein names
- Entities share the same name
- Use of contextual information of known occurrences to learn weights; apply weights to elements of unknown occurrences allows classification
- Terminological information, morphological, POS tags, shallow syntactic info, etc.

# Term structuring

- **term clustering** (linking semantically similar terms) and **term classification** (assigning terms to classes from a pre-defined classification scheme)
- **Hypothesis**: similar terms tend to appear in similar contexts (patterns)
- combining various sources of similarity:
  - lexical
  - syntactic
  - contextual
  - Ontological (using external resources)

# Term structuring

- Based on term similarities
  - choice of features:
    - domain specific → ontology
    - linguistic → text
- ontology-based similarity
- textual similarity
  - internal features
  - contextual features

# Using ontologies

- two terms should match if they are:
    - identified as **variants**
    - **siblings** in the is-a hierarchy
    - in the **is-a** or **part-whole** relation
  - the **distance** between the corresponding nodes in the ontology should be transformed into the matching score
- I. Spasic presentation MIE Tutorial <http://www.nactem.ac.uk/>

# Using text

- number of neologisms: terms are not in the ontologies
- Use of text based techniques to calculate similarities
- **edit distance** (ED) – the minimal number (or cost) of changes needed to transform one string into the other

- **edit operations:**

| insertion          | deletion           | replacement        | transposition      |
|--------------------|--------------------|--------------------|--------------------|
| ... <b>a-c</b> ... | ... <b>abc</b> ... | ... <b>abc</b> ... | ... <b>abc</b> ... |
| ... <b>abc</b> ... | ... <b>a-c</b> ... | ... <b>adc</b> ... | ... <b>acb</b> ... |

- use of **dynamic programming**

# Examples

ED( vitamin A,  
vitamin-A) = 1 (1 replacement)

ED( vitamin A,  
vitamin C) = 1 (1 replacement)

ED( vitamin A,  
A vitamin) = 4 (2 insertions, 2 deletions)

# Term similarities

- **lexical similarity**: based on sharing term head and/or modifier(s) --hyponymy

nuclear receptor

orphan nuclear receptor

- Sharing heads

progesterone receptor oestrogen receptor

- Specific types of associations
  - mainly general *is\_a* and *part\_of*
  - some domain-specific, e.g. *binding*: *CREP binding protein*

# Contextual similarities

- Features from context
  - syntactic category
  - terminological status
  - position relative to the term
  - syntactic relation between a context element and the term
  - semantic properties
  - semantic relation between a context element and the term .....

# Lexical & syntactic patterns

- a lexico-syntactic pattern:  
... **Term** (, **Term**)\* [,] and **other Term** ...
- the leading **Terms** hyponyms of the head **Term**  
... **antiandrogens**, **hydroxyflutamide**, **bicalutamide**,  
**cyproterone acetate**, **RU58841**, and other **compounds** ...
- candidate instances of the hyponymy relation:  
hyponym( **antiandrogens**, **compound** )  
hyponym( **hydroxyflutamide**, **compound** )  
hyponym( **bicalutamide**, **compound** )  
hyponym( **cyproterone acetate**, **compound** )  
hyponym( **RU58841**, **compound** )

# Contextual information

- automatic pattern mining for most important context patterns
  - find most important contexts in which a term appears

... receptor is *bound* to these DNA sequences ...  
 ... proteins *bound* to the DNA ...  
 ... estrogen receptor *bound* to DNA ...  
 ... steroid receptor coactivator-1 when *bound* to DNA ...  
 ... progesterone receptor complexes *bound* to DNA ...  
 ... RXRs *bound* to respective DNA elements in vitro ...  
 ... glucocorticoid receptor to *bind* DNA ...

pattern:                    <TERM> V:*bind* <TERM:DNA>

# Stumbling blocks

- Lexical similarities affected by many neologisms and ad hoc names
  - only 5% of most frequent terms in GENIA belonging to same biomedical class have some lexical links
- Issue over how much context to use (sentence, phrase, abstract, ...)
- Attempts at using co-occurrence: many report up to 40% of co-occurrence based relationships biologically meaningless

# Term similarities

- **SOLD** = **S**yntactic, **O**ntology-driven & **L**exical **D**istance (Spasic, I. & Ananiadou, S. 2005)
- hybrid approach to comparing term contexts, which relies on:
  - **linguistic** information (acquired through tagging and parsing)
  - **domain-specific** knowledge (obtained from the ontology)
- roughly based on the **approximate pattern matching** (i.e. ED)
- combines **ontology-based** similarity with **corpus-based** similarity using both **internal** and **contextual features**

# Term similarity measures

- the ED is used to account for **structural differences** in term contexts while making it **more flexible** with respect to lexical and terminological variations
- **approximate** matching not only for a term context as a **whole**, but for its **individual constituents** as well
- different types of features combined:
  - syntactic
  - lexical
  - semantic

# Context alignment

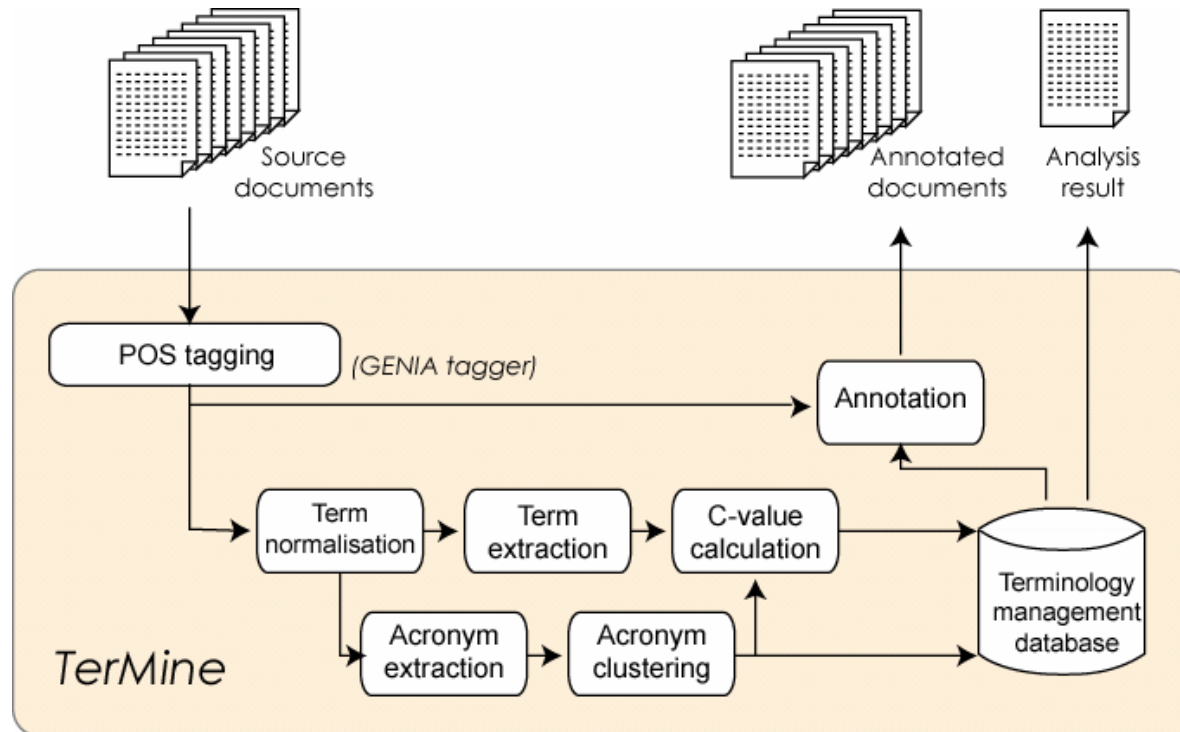
The |-----| **ecdysone receptor** | ( | EcR | ) | is | a  
member |

The | classical | **receptor** | for | estradiol | - | is | a  
member |

of | the | large family | of | nuclear hormone receptors | , | which |  
of | a | super-family | of | nuclear | receptors | - | that |

are | -- | ligand | regulated | transcription factors | .  
function | as | hormone | regulated | transcription factors | .

# A Term management system



<http://www.nactem.ac.uk/>

TerMine [GUI]

File Edit View Help

Source documents

**Beta-arrestin binding to the beta2-adrenergic receptor requires both receptor phosphorylation and receptor activation.**  
Krasel C, Bunemann M, Lorenz K, Lohse MJ.  
Institute for Pharmacology and Toxicology, Versbacher Strasse 9, D-97078 Wurzburg, Germany.

Homologous desensitization of beta2-adrenergic receptors has been shown to be mediated by phosphorylation of the agonist-stimulated receptor by G-protein-coupled receptor kinase 2 (GRK2) followed by binding of beta-arrestins to the phosphorylated receptor. Binding of beta-arrestin to the receptor is a prerequisite for subsequent receptor desensitization, internalization via clathrin-coated pits, and the initiation of alternative signaling pathways. In this study we have investigated the interactions between receptors and beta-arrestin2 in living cells using fluorescence resonance energy transfer. We show that (a) the initial kinetics of beta-arrestin2 binding to the receptor is limited by the kinetics of GRK2-mediated receptor phosphorylation; (b) repeated stimulation leads to the accumulation of GRK2-phosphorylated receptor, which can bind beta-arrestin2 very rapidly; and (c) the interaction of beta-arrestin2 with the receptor depends on the activation of the receptor by agonist because agonist withdrawal leads to swift dissociation of the receptor-beta-arrestin2 complex. This fast agonist-controlled association and dissociation of beta-arrestins from prephosphorylated receptors should permit rapid control of receptor sensitivity in repeatedly stimulated cells such as neurons.

**beta2-adrenergic receptor gene single-nucleotide polymorphisms are associated with rheumatoid arthritis in northern Sweden.**  
Xu B, Arlehaug L, Rantapaa-Dahlquist SB, Lefvert AK.  
Department of Immunology, American Red Cross Biomedical Research and Development, MD 20855, USA. xubiy@usa.redcross.org

The beta2-adrenergic receptor (beta2-AR) belongs to the group of G-protein-coupled receptors and is present on skeletal and cardiac muscle cells and on lymphocytes. The gene encoding beta2-AR (ADRB2) displays a moderate degree of heterogeneity in the human population and the distributions of single-nucleotide polymorphisms (SNPs) at amino acid positions 16, 27, and 164 are changed in asthma, obesity, and hypertension and in the autoimmune disease myasthenia gravis. An involvement of the beta2-AR has also been suggested in human rheumatoid arthritis (RA) and its animal model. We describe here an increased prevalence of the alleles Arg16 and Gln27 and a lower prevalence of homozygosis for Gly16 and Glu27 in patients with RA. Patients having the genotype combination GlyGly16-GlnGlu27 had higher levels of rheumatoid factor (RF) and a more active disease than other patients. Patients having the genotype Arg16-Gln27+ had higher levels of RF when compared to those having Arg16+Gln27+, and patients who were carriers of Gln27 had a more active disease than non-carriers of Gln27. Our results show an association of beta2-AR SNPs with RA in a population from the northern part of Sweden. Our study also confirms the strong linkage disequilibrium of genotypes at amino acid

Result 1 - 50 of about 1131 terms

| Rank | Term                                      | Score   |
|------|---|---------|
| 1    | beta2-adrenergic receptor                 | 65.7778 |
| 2    | blood pressure                            | 16.8    |
| 3    | beta2-adrenergic receptor gene            | 14.8496 |
| 4    | single nucleotide polymorphisms           | 9.50977 |
| 5    | adrenergic receptor                       | 9.14286 |
| 6    | Gly16 allele                              | 8       |
| 7    | A549 cells                                | 8       |
| 8    | body mass index                           | 7.92481 |
| 9    | cystic fibrosis patients                  | 7.92481 |
| 10   | protein kinase                            | 7.625   |
| 11   | cystic fibrosis                           | 7.33333 |
| 12   | confidence interval                       | 7       |
| 13   | metabolic syndrome                        | 7       |
| 14   | allelic frequency                         | 7       |
| 15   | bioluminescence resonance energy transfer | 6.8     |

Ready

# Challenges of biomedical terminology

- Linking termforms in text with existing resources
- Term clustering, classification and linking to databases, ontologies
- Selection of most representative terms (concepts) in documents (important for improved IR, database curation, annotation tasks)

# Information Extraction in Biology

- Results appear depressed compared to general language
  - Dependent of earlier stages of processing (tokenisers, taggers, results from NER, etc)
  - MUC data 80% F-score template relations, 60% events
  - Challenge for biotext mining is to achieve similar results
    - Evaluation see Hirschman, L. (Text mining book)

# IE in Biology

- Pattern-matching
- Context-free grammar approaches
- Full parsing approaches
- Sublanguage driven IE
- Ontology-driven IE

McNaught, J. & Black, W. (2006) Information Extraction, [Text Mining for Biology & Biomedicine](#), Artech house, pp.143-177

# Pattern-matching IE

- Usual limitations with non inclusion of semantic processing
- Large amount of surface grammatical structures = too many patterns (Zipf's law)
- Cannot explore syntactic generalisations (active, passive voice)
- Systems extract phrases or entire sentences with matched patterns; restricted usefulness for subsequent mining

# Pattern-matching systems (1)

- BioIE uses patterns to extract sentences, protein families, structures, functions..
  - Presents user with relevant information, improvement from classic IR
- BioRAT uses “deeper” analysis, tagging, apply RE over POS tags, stemming, gazetter categories etc
  - Templates apply to extract matching phrases, primitive filters (verbs are not proteins, etc)

# Pattern matching systems (2)

- RLIMS-P (Hu) protein phosphorylation by looking for enzymes, substrates, sites assigned to *agent*, *theme*, *site* roles of phosphorylation relations
- Pos tagger, trained on newswire, chunking, semantic typing of chunks, identification of relations using pattern-matching rules
- Semantic typing of NPs: using combination of clue words, suffixes, acronyms etc
- Semantically typed sentences matched with rules
- Patterns target sentences containing *phosphorylate*

# CFG approaches

- Application of CFG; little consideration of linguistic constraints
- Temkin & Gilder (2003) use compiler generator tools to produce lexical analyser and parser to extract gene, protein, molecule interactions
- Top-down parser, hypotheses made about structure of sentence before seeing the words
- Rely on small number of interaction keywords, grouped by semantic category
- Problems with long complex sentences because grammar is simple

# Full parsing approaches

- Link Grammar applied for protein-protein interactions; general English grammar adapted to bio-text
- Link Grammar finds all possible *linkages* according to its grammar
- Number of analyses reduced by random sampling, heuristics, processing constraints relaxed
  - 10,000 results permitted per sentence
  - 60% of protein interactions extracted
  - Problems: missing possessive markers & determiners, coordination of compound noun modifiers

# Full parsing IE (2)

- Not all parsing strategies suitable for bio-text mining
- Text type, abstracts, “ungrammaticality” related with sublanguage characteristics?
- Ambiguity and full parsing; fragmentary phrases (titles, headings, text in table cells, etc)
- **CADERIGE** project used Link grammar but on shallow parsing mode
- **Kim & Park (BioIE)** use combinatorial categorial grammar, annotated with GO concepts, extract general biological interactions
- 1,300 patterns applied to find instances of patterns with keywords

# Full parsing (3)

- Keywords indicate basic biological interactions
- Patterns find potential arguments of the interaction keywords (verbs or nominalisations)
  - Validated arguments mapped into GO concepts
  - Difficult to generalise interaction keyword patterns
- BioIE's syntactic parsing performance improved after adding subcategorisation frames on verbal interaction keywords

# Full parsing (4)

- Daraselia(2004) use full parsing and domain specific filter to extract protein interactions
  1. All syntactic analyses discovered using CFG and variant of LFG
  2. Each alternative parse mapped to its corresponding semantic representation
  3. Output= set of semantic trees, lexemes linked by relations indicating thematic or attributive roles
  4. Apply custom-built, frame based ontology to filter representations of each sentence
  5. Preference mechanism controls construction of frame tree, high precision, low recall (21%)

# Sublanguage-driven IE (1)

- Language of a special community (e.g. biology)
- Particular set of constraints re GL
- Constraints operate at all linguistic levels
  - Special vocabulary (terms)
  - Specialised term formation rules
  - Sublanguage syntactic patterns
  - Sublanguage semantics
- These constraints give rise to the *informational structure* of the domain (Z. Harris)
- See JBI 35(4) Special Issue on Sublanguage

# GENIES system

- Employs SL approach to extract biomolecular interactions
- Uses hybrid syntactic-semantic rules
  - Syntactic and semantic constraints referred to in one rule
- Able to cope with complex sentences
- Frame-based representation
  - Embedded frames
- Domain specific ontology covers both entities and events

# GENIES system

- Default strategy: full parsing
  - Robust due to sublanguage constraints
  - Much ambiguity excluded
- If full parse fails, partial parsing invoked
  - Maintains good level of recall
- Precision: 96%, Recall: 63%

# Ontology-driven IE

- Until recently most rule based IE have used neither linguistic lexica nor ontologies
  - Reliance on gazetteers
  - Small number of semantic categories
- Gazetteer approach not well suited in bioIE
- **Ontology based** vs **ontology driven**
  - Passive use of ontologies, map discovered entity to concept
  - Active use, ontology guides and constrains analysis, fewer rules
- Examples: PASTA, GenIE not SL
- GENIES, SL and ontology driven

# Summary: simple pattern matching

- Over text strings
  - Many patterns required, no generalisation possible
- Over POS
  - Some generalisation but ignore sentence structure
- POS tagging, chunking, semantic p-m, typing
  - Limited generalisation, some account taken of structure, limited consideration of SL patterns

# Summary: full parsing

- Full parsing on its own, parsing done in combination with chunking, partial parsing, heuristics) to reduce ambiguity, filter out implausible readings
  - GL theories not appropriate
  - Difficult to specialise for biotext
  - Many analyses per sentence
  - Missing information due to sublanguage meaning

# Summary: sublanguage approach

- Exploits a rich SL lexicon
- Describes SL verbs in detail
- Syntactic-semantic grammar
- Current systems would benefit from adopting ontology-driven approach

# Ontology-driven

- Uses event concept frames to guide processing
- Integration of extracted information
- Current systems would benefit from adopting also SL approach

# Linguistically Annotated Corpora

- GENIA
  - Domain
    - Mesh term: *Human, Blood Cells, and Transcription Factors.*
  - Annotation: POS, named entity, parse tree
- Penn BioIE
  - Domain
    - the molecular genetics of oncology
    - the inhibition of enzymes of the CYP450 class.
  - Annotation: POS, named entity, parse tree
- Yapex
- GENETag
- etc..

# Part-Of-Speech annotation

The peri-kappa B site mediates human immunodeficiency  
 DT NN NN NN VBZ JJ NN  
 virus type 2 enhancer activation in monocytes ...  
 NN NN CD NN NN IN NNS

| Corpus     | size            |
|------------|-----------------|
| GENIA      | 2,000 abstracts |
| Penn BioIE | 2,157 abstracts |
| MedPost    | 5,700 sentences |

# Named-entity annotation

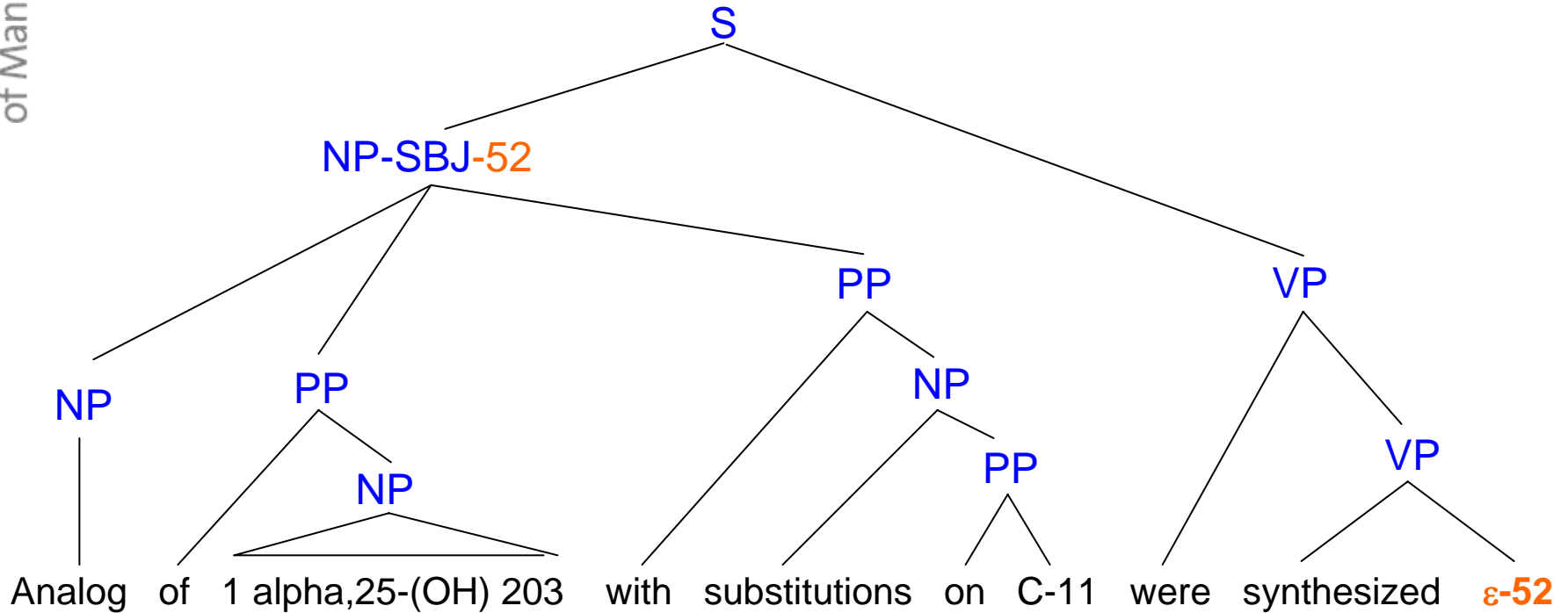
We have shown that **interleukin-1 (IL-1)** and **IL-2** control  
**protein protein protein**  
**IL-2 receptor alpha (IL-2R alpha) gene** transcription in  
**DNA**  
**CD4-CD8-murine T lymphocyte precursors.**  
**cell\_line**

- Entity type
  - Genes/protein names
  - Enzymes, substances, etc.
  - GENIA ontology

# Entity-annotated corpora (Kim, 2006)

| Corpus                      | Annotation Type                     | Remarks  |
|-----------------------------|-------------------------------------|--|
| GENIA<br>GENIA-JNLPBA       | (Size)<br>Term (2,000<br>abstracts) | Terminal concepts in<br>GENIA ontology<br>Proteins, DNAs,<br>RNAs, cell lines, cell<br>types |
| PennBioIE-Oncology          | Entity (1,157<br>abstracts)         | Genes, variation<br>events, malignancies   |
| PennBioIE-CYP               | Entity (1,100<br>abstracts)         | CYP450 enzymes,<br>other substances,<br>quantitative<br>measurements                         |
| GENETAG-05                  | Entity (15,000<br>sentences)        | Gene/protein names   |
| (MedTag)<br>ABGene (MedTag) | Entity (4,265<br>sentences)         | Gene/protein names   |
| Yapex                       | Entity (200 abstracts)              | Protein names  |

# Phrase-structure annotation



| Corpus     | Size           |
|------------|----------------|
| GENIA      | 1500 abstracts |
| Penn BioIE | 642 abstracts  |

# Other types of annotation

- Co-reference
- Biological events
  - “Lipopolysaccharide induces phosphorylation of MAD3”
    - Event #1
      - Type : Protein\_amino\_acid\_phosphorylation (GO:0006468)
      - Theme: *MAD3* (Protein\_molecule)
    - Event #2
      - Type : Positive\_regulation (GO:0048518)
      - Theme: Event #1
      - Cause : *Lipopolysaccharide*

# Basic Steps of NLP

- Sentence splitting
- Tokenization
- Part-of-speech tagging
- Shallow parsing (chunking)
- Named entity recognition
- CFG parsing
- Deep parsing

# Sentence splitting

- PubMed has no information about sentence boundaries.
- Sentence splitting by simple heuristic rules
  - Space + capital letter
  - Exceptions: “Dr. Xxx”, “e.g. YYY”, ...
- JASMINE
  - A rule-based sentence splitter
  - [http://uvdb3.hgc.jp/ALICE/program\\_download.html](http://uvdb3.hgc.jp/ALICE/program_download.html)
- Machine learning
  - Maximum-entropy (Reynar, 1997): 98-99% accuracy
  - Training data: e.g. GENIA

# Tokenization

- Convert a sentence into a sequence of *tokens*
  - tokenizer.sed: a simple sed script
    - <http://www.cis.upenn.edu/~treebank/tokenization.html>
- Undesirable tokenization
  - org: “1,25(OH)2D3”
  - tokenized: “1 , 25 ( OH ) 2D3”
- Tokenization for biomedical text
  - Not straight-forward
  - Needs dictionary? Machine learning?

# Part-Of-Speech tagging

The peri-kappa B site mediates human immunodeficiency  
 DT NN NN NN VBZ JJ NN  
 virus type 2 enhancer activation in monocytes ...  
 NN NN CD NN NN IN NNS

- Assign a part-of-speech tag to each token in a sentence.

# POS tagging algorithms

- Accuracies on the WSJ corpus

|                              | <b>Training<br/>Cost</b> | <b>Accura<br/>cy</b> |
|------------------------------|--------------------------|----------------------|
| Dependency Net (2003)        |                          | 97.2                 |
| Perceptron (2002)            |                          | 97.1                 |
| SVM (2003)                   |                          | 97.1                 |
| Bidirectional MEMM<br>(2005) |                          | 97.1                 |
| Brill's tagger (1995)        | low                      | 96.6                 |
| HMM (2000)                   | low                      | 96.7                 |

# POS taggers

- Brill's tagger
  - <http://www.cs.jhu.edu/~brill/>
- TnT tagger
  - <http://www.coli.uni-saarland.de/~thorsten/tnt/>
- Stanford tagger
  - <http://nlp.stanford.edu/software/tagger.shtml>
- SVMTool
  - <http://www.lsi.upc.es/~nlp/SVMTool/>
- GENIA tagger
  - <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/>

# Tagging errors made by a WSJ-trained POS tagger

... and membrane potential after mitogen ~~binding~~.

CC NN NN IN NN JJ

... two factors, which ~~bind~~ to the same kappa B enhancers...

CD NNS WDT NN TO DT JJ NN NN NNS

... by analysing the ~~Ag~~ amino acid sequence.

IN VBG DT VBG JJ NN NN

... to contain ~~more~~ T-~~cell~~ determinants than ...

TO VB RBR JJ NNS IN

Stimulation of interferon beta gene transcription ~~in vitro~~ by

NN IN JJ JJ NN NN IN NN IN

# Taggers for general text do not work well on biomedical text

Performance of the Brill tagger evaluated on randomly selected 1000 MEDLINE sentences: 86.8% (Smith et al., 2004)

|                      | Accuracy |
|----------------------|----------|
| Exact                | 84.4%    |
| NNP = NN, NNPS = NNS | 90.0%    |
| LS = NN              | 91.3%    |
| JJ = NN              | 94.9%    |

Accuracies of a WSJ-trained POS tagger evaluated on the GENIA corpus (Tsuruoka et al., 2005)

# MedPost

(Smith et al., 2004)

- Hidden Markov Models (HMMs)
- Training data
  - 5700 sentences randomly selected from various thematic subsets.
- Accuracy
  - 97.43% (native tagset), 96.9% (Penn tagset)
  - Evaluated on 1,000 sentences
- Available from
  - <ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz>

# Training POS taggers with bio-corpora

(Tsuruoka and Tsujii, 2005)

| training \              | WSJ  | GENIA | PennBioIE |
|-------------------------|------|-------|-----------|
| WSJ                     | 97.2 | 91.6  | 90.5      |
| GENIA                   | 85.3 | 98.6  | 92.2      |
| PennBioIE               | 87.4 | 93.4  | 97.9      |
| WSJ + GENIA             | 97.2 | 98.5  | 93.6      |
| WSJ + PennBioIE         | 97.2 | 94.0  | 98.0      |
| GENIA + PennBioIE       | 88.3 | 98.4  | 97.8      |
| WSJ + GENIA + PennBioIE | 97.2 | 98.4  | 97.9      |

# How robust are they?

Relative performance evaluated on recent abstracts selected from three journals:

- Nucleic Acid Research (NAR)
- Nature Medicine (NMED)
- Journal of Clinical Investigation (JCI)

| training                   | NA  | NME | NME | Total (Acc.) |
|----------------------------|-----|-----|-----|--------------|
| WSJ                        | 109 | 47  | 102 | 258 (70.9%)  |
| GENIA                      | 121 | 74  | 132 | 327 (89.8%)  |
| PennBioIE                  | 129 | 65  | 122 | 316 (86.6%)  |
| WSJ + GENIA                | 125 | 74  | 135 | 334 (91.8%)  |
| WSJ + PennBioIE            | 133 | 71  | 133 | 337 (92.6%)  |
| GENIA + PennBioIE          | 128 | 75  | 135 | 338 (92.9%)  |
| WSJ + GENIA +<br>PennBioIE | 133 | 74  | 139 | 346 (95.1%)  |

PennBioIE

# Chunking (shallow parsing)

He reckons the current account deficit will narrow to  
 NP VP NP VP PP  
only # 1.8 billion in September.  
 NP PP NP

- A chunker (shallow parser) segments a sentence into non-recursive phrases.

# Extracting noun phrases from MEDLINE

(Bennett, 1999)

- Rule-based noun phrase extraction
  - Tokenization
  - Part-Of-Speech tagging
  - Pattern matching

Noun phrase extraction accuracies evaluated on 40 abstracts

|          | FastNPE | NPtool | Chopper | AZ<br>Phraser |
|----------|---------|--------|---------|---------------|
| Recall   | 50%     | 95%    | 97%     | 92%           |
| Precisio | 80%     | 96%    | 90%     | 86%           |

n

# Chunking with Machine learning

- Chunking performance on Penn Treebank

|  | Recall | Precisi | F-    |
|--|--------|---------|-------|
| Winnow (with basic features) (Zhang, 2002) | 93.60  | 93.54   | 93.57 |
| Perceptron (Carreras, 2003)                | 93.29  | 94.19   | 93.74 |
| SVM + voting (Kudoh, 2003)                 | 93.92  | 93.89   | 93.91 |
| SVM (Kudo, 2000)                           | 93.51  | 93.45   | 93.48 |
| Bidirectional MEMM (Tsuruoka, 2005)        | 93.70  | 93.70   | 93.70 |

# Machine learning-based chunking

- Convert a treebank into sentences that are annotated with chunk information.
  - CoNLL-2000 data set
    - <http://www.cnts.ua.ac.be/conll2000/chunking/>
    - The conversion script is available
- Apply a sequence tagging algorithm such as HMM, MEMM, CRF, or Semi-CRF.
- YamCha: an SVM-based chunker
  - <http://www.chasen.org/~taku/software/yamcha/>

# GENIA tagger

- Algorithm: Bidirectional MEMM
- POS tagging
  - Trained on WSJ, GENIA and Penn BioIE
  - Accuracy: 97-98%
- Shallow parsing
  - Trained on WSJ and GENIA
  - Accuracy: 90-94%
- Can output base forms
- Available from
  - <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/>

# Named-Entity Recognition

We have shown that **interleukin-1 (IL-1)** and **IL-2** control  
**protein protein protein**  
**IL-2 receptor alpha (IL-2R alpha) gene** transcription in  
**DNA**  
**CD4-CD8-murine T lymphocyte precursors.**  
**cell\_line**

- Recognize named-entities in a sentence.
  - Gene/protein names
  - Protein, DNA, RNA, cell\_line, cell\_type

# Performance of biomedical NE recognition

- Shared task data for Coling 2004 BioNLP workshop
  - entity types: protein, DNA, RNA, cell\_type, and cell\_line

|                             | <b>Recall</b> | <b>Precisi</b> | <b>F-</b> |
|-----------------------------|---------------|----------------|-----------|
| SVM+HMM (Zhou, 2004)        | 76.0          | 69.4           | 72.6      |
| Semi-Markov CRFs (in prep.) | 72.7          | 70.4           | 71.5      |
| Two-Phase (Kim, 2005)       | 72.8          | 69.7           | 71.2      |
| Sliding Window (in prep.)   | 71.5          | 70.2           | 70.8      |
| CRF (Settles, 2005)         | 72.0          | 69.1           | 70.5      |
| MEMM (Finkel, 2004)         | 71.6          | 68.6           | 70.1      |
| :                           | :             | :              | :         |

# Features

Classification models, main features used in NLPBA (Kim, 2004)

|     | CM | lx | af | or | sh | g                 | gz  | po | n | sy | tr  | a                 | ca | do | p | pr | ext.                |
|-----|----|----|----|----|----|-------------------|-----|----|---|----|-----|-------------------|----|----|---|----|---------------------|
| Zho | SH |    | x  | x  |    | <del>n</del><br>x | x   | x  | p |    | x   | <del>b</del><br>x | x  |    | a | x  |                     |
| Fin | M  | x  | x  |    | x  |                   | x   | x  |   | x  |     | x                 |    | x  | x | x  | B,                  |
| Set | C  | x  | x  | x  | x  |                   | (x) |    |   |    | (x) |                   |    |    |   | x  | <del>W</del><br>(W) |
| Son | SC |    | x  | x  |    |                   |     | x  | x |    |     |                   |    |    |   | x  | V                   |
| Zha | H  | x  |    |    |    |                   |     |    |   |    |     |                   |    |    |   | x  | M                   |

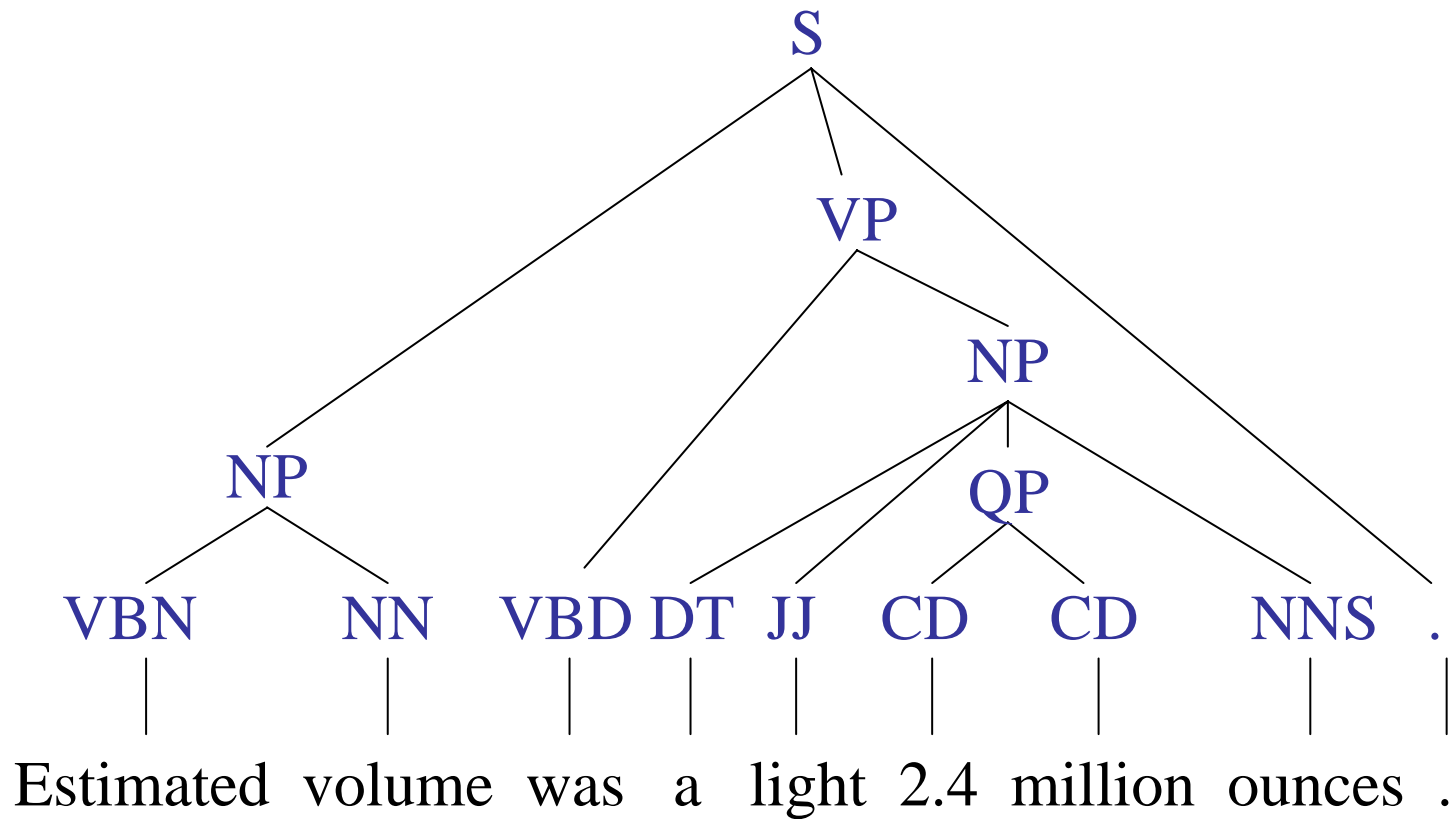
Classification Model (CM):

S: SVM; H: HMM; M: MEMM; C: CRF

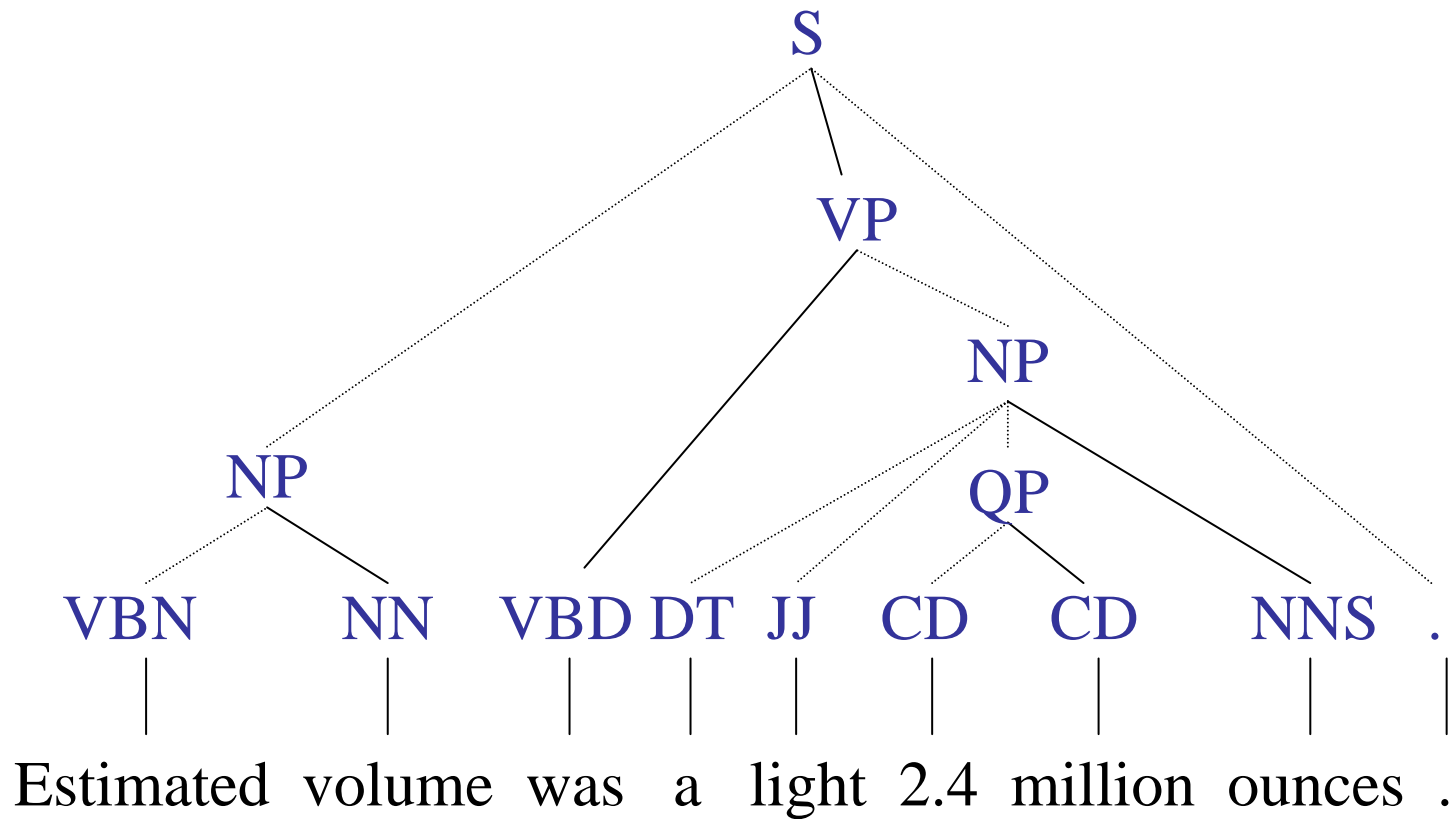
Features

lx: lexical features; af: affix information (chracter n-grams); or; orthographic Information; sh: word shapes; gn: gene sequence; gz: gazetteers; po: part-of-speech tags; np: noun phrase tags; sy: syntactic tags; tr: word triggers; ab: abbreviations; ca: cascaded entities; do: global document information; pa: parentheses handling; pre: previously predicted entity tags; B: British National Corpus; W: WWW; V: virtually generated corpus; M: MEDLINE

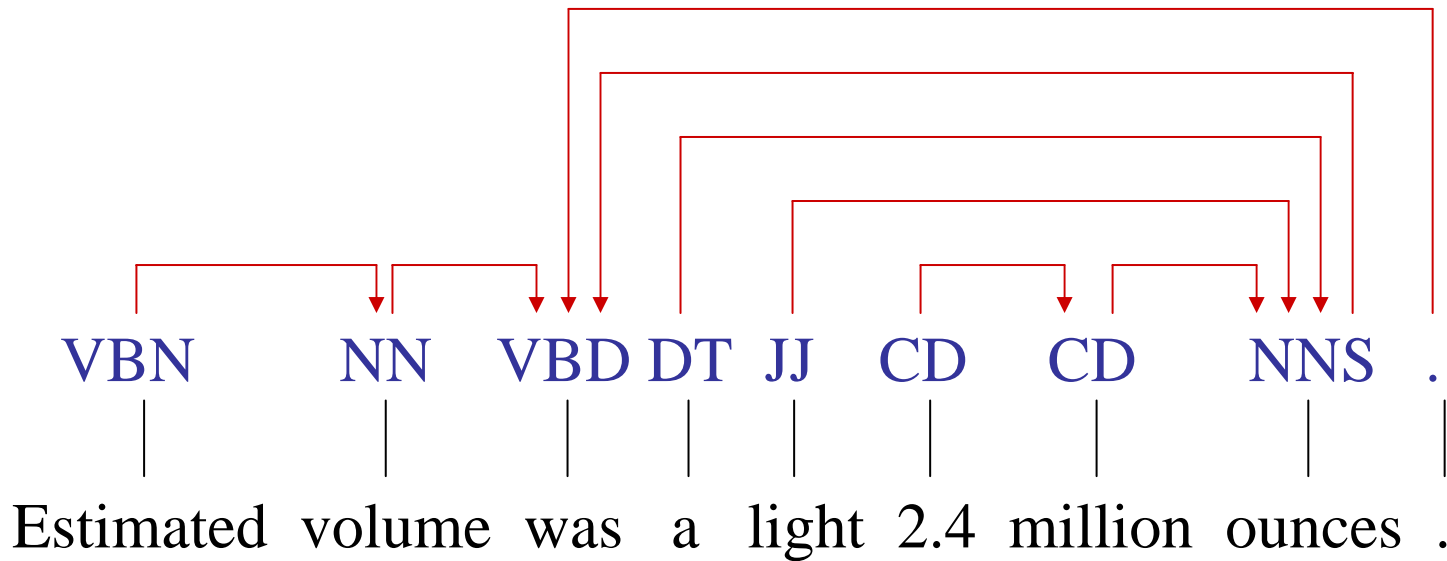
# CFG parsing



# Phrase structure + head information



# Dependency relations



# CFG parsing algorithms

- Performance on the Penn Treebank

|  | LR   | LP   | F-                              |
|--|------|------|---------------------------------|
| Generative model (Collins, 1999)               | 88.1 | 88.3 | <del>88.2</del><br><b>score</b> |
| Maxent-inspired (Charniak, 2000)               | 89.6 | 89.5 | 89.5                            |
| Simply Synchrony Networks (Henderson,<br>2004) | 89.8 | 90.4 | 90.1                            |
| Data Oriented Parsing (Bod, 2003)              | 90.8 | 90.7 | 90.7                            |
| Re-ranking (Johnson, 2005)                     |      |      | 91.0                            |

# CFG parsers

- Collins parser
  - <http://people.csail.mit.edu/mcollins/code.html>
- Bikel's parser
  - <http://www.cis.upenn.edu/~dbikel/software.html#stat-parser>
- Charniak parser
  - <http://www.cs.brown.edu/people/ec/>
- Reranking parser
  - <http://www.cog.brown.edu:16080/~mj/Software.htm>
- SSN parser
  - [http://homepages.inf.ed.ac.uk/jhender6/parser/ssn\\_parser.htm](http://homepages.inf.ed.ac.uk/jhender6/parser/ssn_parser.htm)  
1

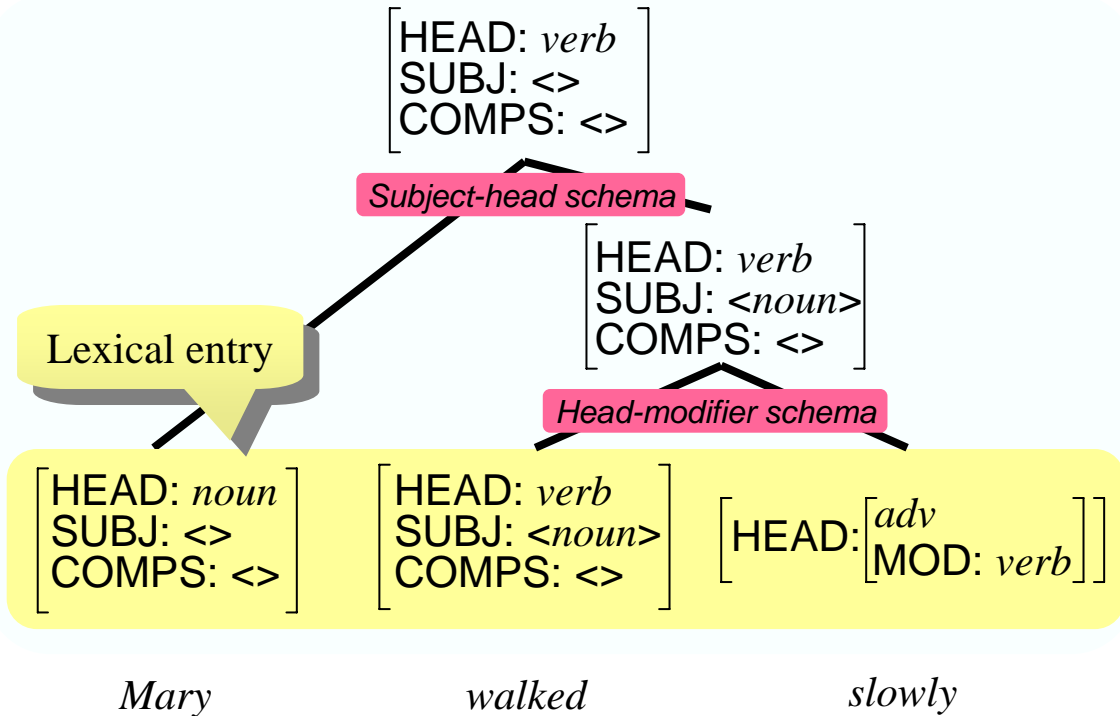
# Parsing biomedical documents

- CFG parsing accuracies on the GENIA treebank (Clegg, 2005)

|                 | LR    | LP    | F-                        |
|-----------------|-------|-------|---------------------------|
| Bikel 0.9.8     | 77.43 | 81.33 | <del>79.33</del><br>score |
| Charniak        | 76.05 | 77.12 | 76.58                     |
| Collins model 2 | 74.49 | 81.30 | 77.75                     |

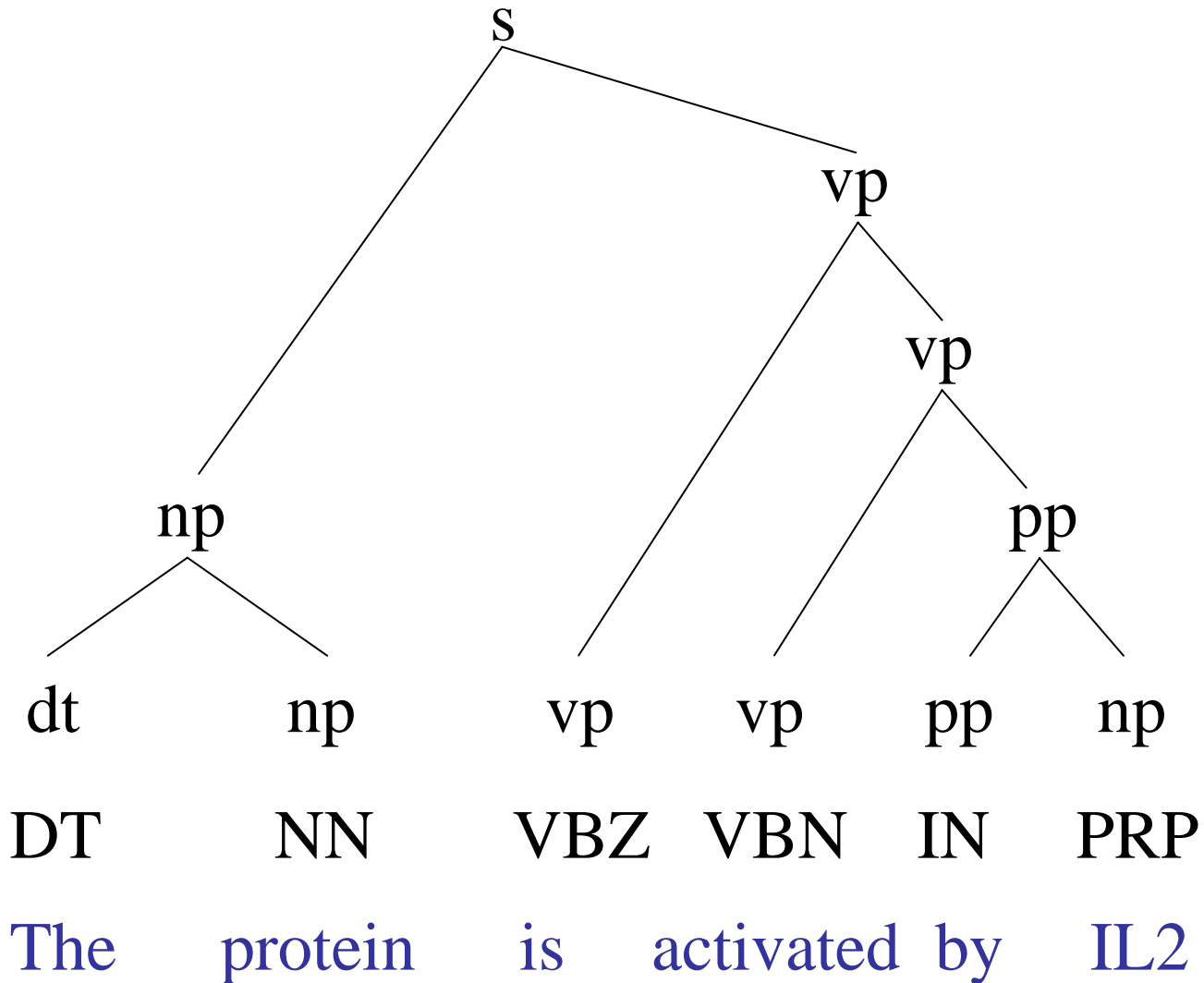
- In order to improve performance,
  - Unsupervised parse combination (Clegg, 2005)
  - Use lexical information (Lease, 2005)
    - 14.2% reduction in error.

# HPSG parsing

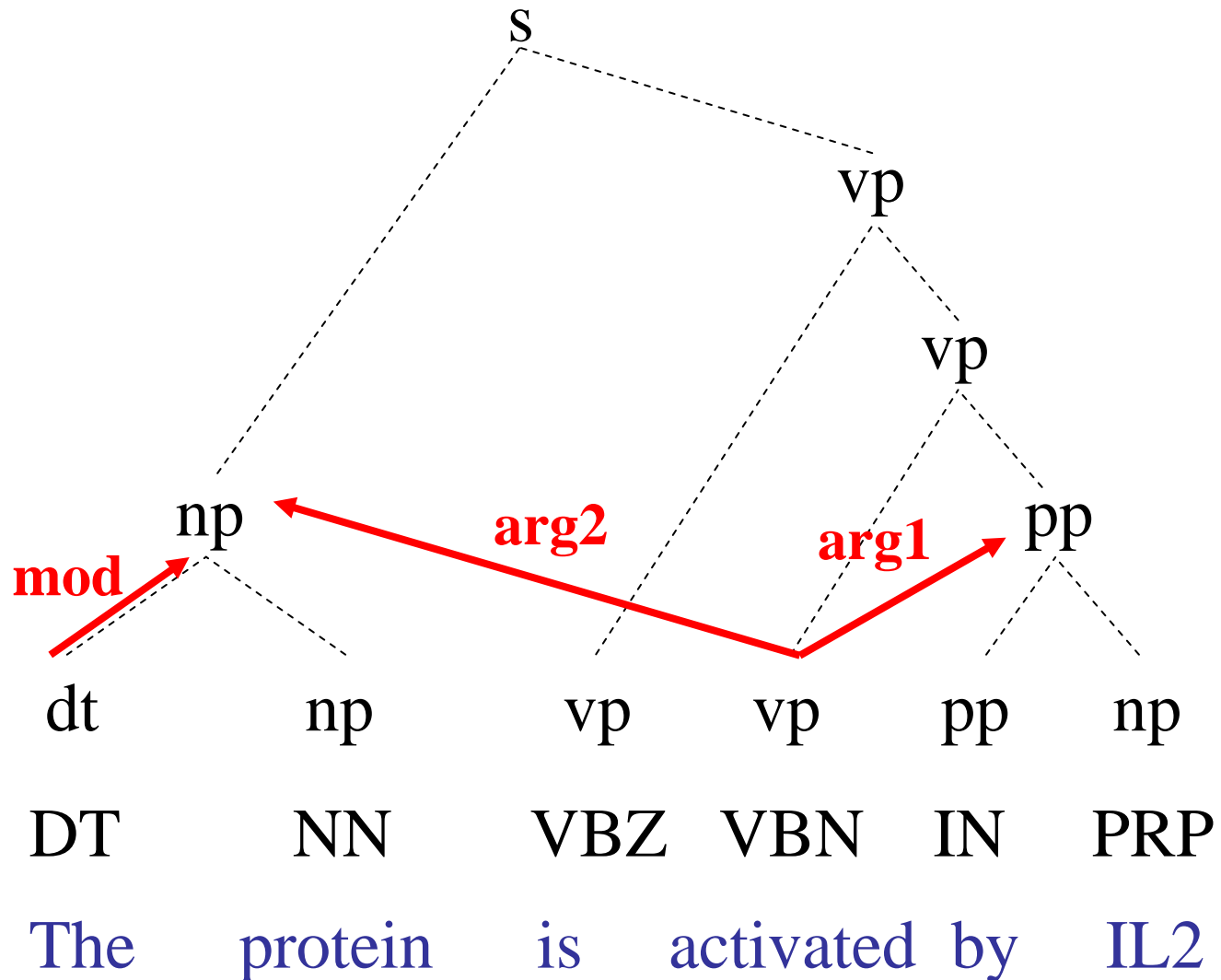


- HPSG
  - A few schema
  - Many lexical entries
  - Deep syntactic analysis
- Grammar
  - Corpus-based grammar construction (Miyao et al 2004)
- Parser
  - Beam search (Tsuruoka et al.)

# Phrase structure



# Predicate-argument relations



# Parsing MEDLINE with HPSG

- Enju
  - A wide-coverage HPSG parser
  - <http://www-tsujii.is.s.u-tokyo.ac.jp/enju/>
- Parsing performance on the GENIA Treebank (Hara et al., 2005)
  - with gold-standard POS tags: 85.1 f-score.
  - Use the GENIA in training: 86.9 f-score

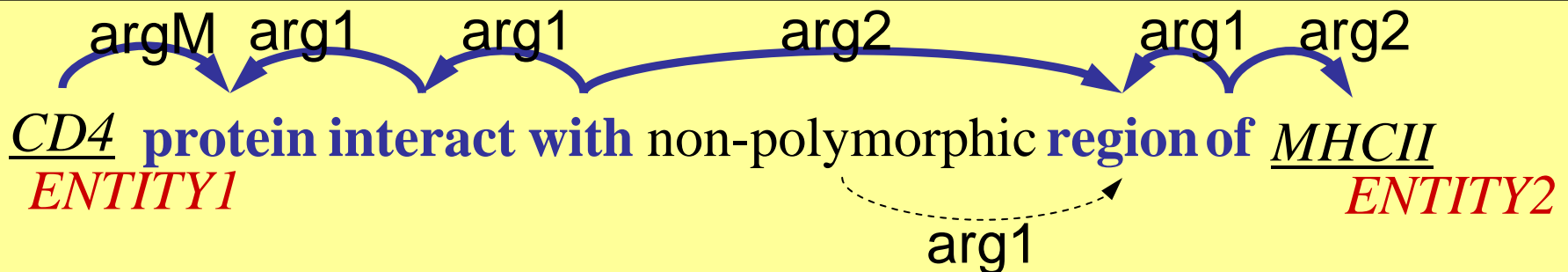
# Extraction of Protein-protein Interactions:

## Predicate-argument relations + SVM

- (Yakushiji, 2005) (1)

CD4 protein interacts with non-polymorphic regions of MHCII .  
*ENTITY1* *ENTITY2*

Extraction patterns based on predicate-argument relations

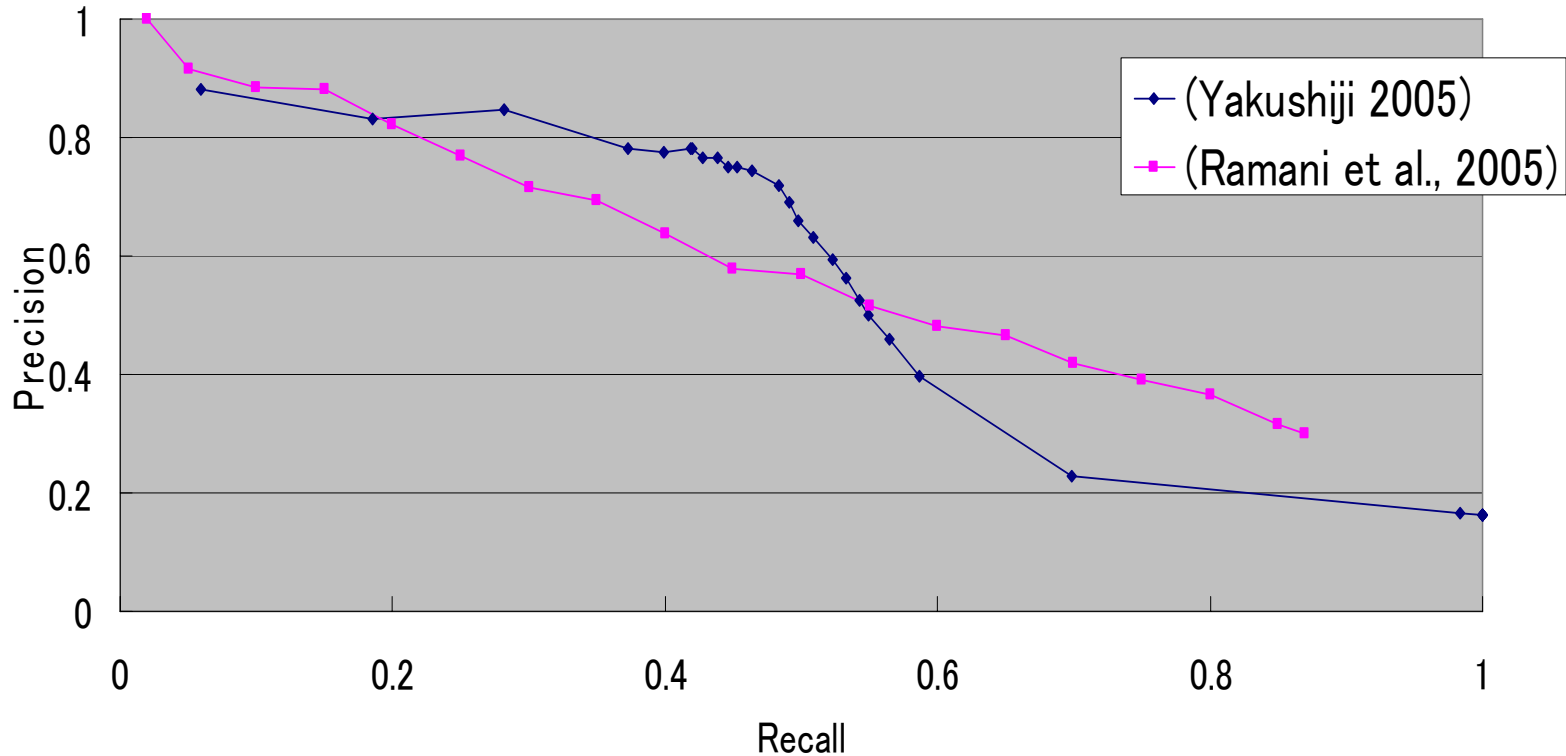


SVM learning with predicate-argument patterns

# Extraction of Protein-protein Interactions:

## Predicate-argument relations + SVM

- Evaluation on the *Aimed* corpus (Bunescu et al., 2004).



# Biomedical IE/IR Systems

- iHOP
  - <http://www.ihop-net.org/UniPub/iHOP/>
- EBIMed
  - <http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp>
- GoPubMed
  - <http://www.gopubmed.org/>
- PubFinder
  - <http://www.glycosciences.de/tools/PubFinder>
- Textpresso
  - <http://www.textpresso.org/>

# MEDUSA

- An interactive IR system based on predicate-argument relations
- System components
  - GENIA tagger
  - Enju (HPSG parser)
  - Dictionary-based named entity recognition
  - IR engine based on region algebra
- (demo)

# MEDUSA

- Subject: p53
- Verb: activate
- Object:

MEDUSA — See what causes cancer?  
MEDUSA is presented by [Tsuji Laboratory](#)

Semantic Search    **Keyword Search**    GCL Search    Custom Search    [User Profile](#)

subject    verb    object

                   [Help](#)

»[Keyword list](#)  
»[Advanced search](#)

Results 1-50 for **p53 activate**    »[Show next](#)    »[Show query](#)    1.52 seconds (6.69% finished)

1.  [PMID: 11779085](#) »XML  
The distribution of **alpha** ( v ) **beta3** is highly restricted , with expression on activated endothelium , activated vascular smooth muscle , tumors , and osteoclasts .
2.  [PMID: 11779500](#) »XML  
Acetylation of **p53** activates transcription through recruitment of coactivators/histone acetyltransferases .
3.  [PMID: 11803461](#) »XML  
Tumor-derived **p53** mutants activate transcription from promoters of various growth-related genes .
4.  [PMID: 11812429](#) »XML  
These changes may be due , at least in part , to induction of **p53** , which activates genes involved in both cell cycle arrest and apoptosis .
5.  [PMID: 11850816](#) »XML  
Electrophoretic mobility shift assays demonstrated that this sequence also is capable of mediating sequence specific binding to p53. **p53** effectively activated transcription through both human and murine **bax gene**

ページが表示されました    インターネット

EA

# Info-PubMed

- An interactive IE system that helps the user to build gene interaction networks.
- System components
  - MEDUSA
  - Extraction of protein-protein interactions
  - Multi-window interface on a browser
- (demo)

# Info-PubMed

Info-PubMed - Microsoft Internet Explorer

ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

戻る 検索 お気に入り

アドレス http://www.tsujii.is.s.u-tokyo.ac.jp/info-pubmed/

移動 リンク Norton AntiVirus

**Gene Searcher**

Search for genes or gene product names

raf1 Search

>> Organism >> Field

**Content Viewer**

INTER ACTION 4 / 2 / 27 NE RAF1 NE Mapk1

NE RAF1 : NE Mapk1 :

Sentences 1 -- 27 Next

- INTER ACTION

PMID15349122 NE RAF1 NE Mapk1

Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- INTER ACTION

PMID15349122 NE RAF1 NE Mapk1

Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- INTER ACTION

PMID15349122 NE RAF1 NE Mapk1

Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- INTER ACTION

PMID15349122 NE RAF1 NE Mapk1

Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- CONDICION PREVENTION

PMID15208680 NE RAF1 NE Mapk1

**B-RAE** depletion inhibits DNA synthesis and induces apoptosis in three melanoma cell lines and we show that the **RAE** inhibitor **BAY43-9006** also blocks **ERK** activity, inhibits DNA synthesis and induces cell death in these cells.
- CONDICION AFFECTION

PMID15208680 NE RAF1 NE Mapk1

**B-RAE** depletion by siRNA blocks **ERK** activity, whereas **A-RAE** and **C-RAE** depletion do not affect **ERK** signalling.
- CONDICION REVERSE

PMID15313890 NE RAF1 NE Mapk1

We found that **SPRY2**, an inhibitor homologous to **SPRY1**, which was previously shown to suppress Ras/**ERK** signaling via direct binding to **Raf-1**, had reduced expression in WT BRAF cells.

**Product**

Oncogene **RAF1**  
**Raf-1**  
v-**raf-1** murine leukemia viral oncogene homolog 1

**A-raf-1**

**Raf-1**  
protein kinase **raf 1**  
**Raf-1** kinase inhibitor protein

**Raf-1**  
v-**raf-1** murine leukemia viral oncogene homolog 1

ページが表示されました

インターネット