

semantic interoperability and data mining in biomedicine  
SEMANTIC MINING

Information Society  
Technologies

# Text Mining

Tutorial at the 2005 NoE "Semantic Mining" Summer School  
Tihany, Hungary, July 2, 2005

**Udo Hahn & Michael Poprat**

FRIEDRICH-SCHILLER-UNIVERSITÄT  
JENA  
[www.coling.uni-jena.de](http://www.coling.uni-jena.de)

ena  
niversity  
anguage  
and  
nformation  
engineering

---

---

---

---

---

---

---

---

## Tutorial Outline

- What is Text Mining?
- Naïve Approach to Text Mining
- Linguistic Approach to Text Mining
- Empirical Approach to Text Mining
- Resources for Empirical NLP
- Summary and Outlook

---

---

---

---

---

---

---

---

## Tutorial Outline

- What is Text Mining?
- Naïve Approach to Text Mining
- Linguistic Approach to Text Mining
- Empirical Approach to Text Mining
- Resources for Empirical NLP
- Summary and Outlook

---

---

---

---

---

---

---

---

## Why NLP for the Life Sciences?

- **Verbal (textual) communication is the primary mode of information exchange**
  - neither biology nor medicine are formal sciences
  - written literature embodies what is known
  - (annotations in) databases are crucial but are manually derived from the literature in whole or in part
- **Literature output grows at very high rates**
  - estimates run on the order of 1 terabyte/week
- **New knowledge is constantly being discovered**
  - entities (concepts), relations (propositions), mechanisms (rules, procedures, plausibility/causality)

---

---

---

---

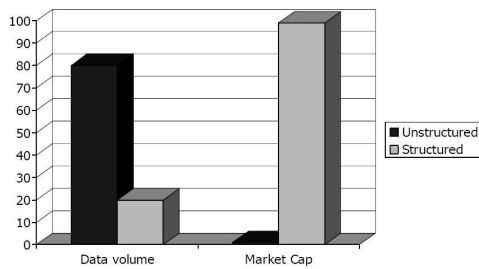
---

---

---

---

## Can BioMed People Afford This?



Source: Prabhakar & Raghavan, Verity (2002)

---

---

---

---

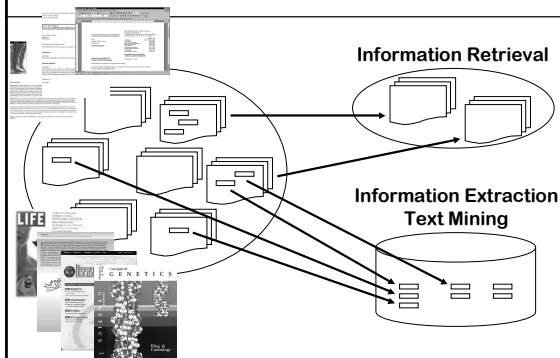
---

---

---

---

## Two Text Analysis Paradigms



---

---

---

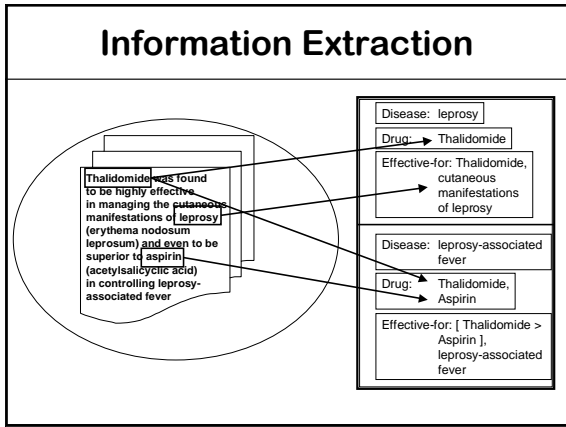
---

---

---

---

---




---

---

---

---

---

---

---

---

### Text Mining as ...

ENTITY MINING		ENTITY MINING
ENTITY 1	INTERACTS-WITH	???

Electrophoretic mobility shift assays indicate that *MS-2beta* and *MS-2gamma* bind to nuclear factors that are induced during U937 differentiation.  
**[bind to , MS-2beta & MS-2gamma , ???]**

... synergistic interactions between overlapping binding sites for the *serum response factor* and *ELK-1* proteins ...  
**[interact with , serum response factor , ???]**

---

---

---

---

---

---

---

---

### Text Mining as Entity Mining

ENTITY MINING		ENTITY MINING
ENTITY 1	INTERACTS-WITH	ENTITY 2

Electrophoretic mobility shift assays indicate that *MS-2beta* and *MS-2gamma* bind to nuclear factors that are induced during U937 differentiation.  
**[bind to , MS-2beta & MS-2gamma , nuclear factors]**

... synergistic interactions between overlapping binding sites for the *serum response factor* and *ELK-1* proteins ...  
**[interact with , serum response factor , ELK-1]**

---

---

---

---

---

---

---

---

### Text Mining as ...

<i>ENTITY 1</i>	RELATION MINING ???	<i>ENTITY 2</i>
-----------------	------------------------	-----------------

Electrophoretic mobility shift assays indicate that *MS-2beta* and *MS-2gamma* bind to nuclear factors that are induced during U937 differentiation.  
[???, *MS-2beta* & *MS-2gamma*, nuclear factors]

... synergistic interactions between overlapping binding sites for the serum response factor and *ELK-1* proteins ...  
[???, serum response factor, *ELK-1*]

---

---

---

---

---

---

---

---

### Text Mining as Relation Mining

<i>ENTITY 1</i>	RELATION MINING INTERACTS-WITH	<i>ENTITY 2</i>
-----------------	-----------------------------------	-----------------

Electrophoretic mobility shift assays indicate that *MS-2beta* and *MS-2gamma* bind to nuclear factors that are induced during U937 differentiation.  
[bind to, *MS-2beta* & *MS-2gamma*, nuclear factors]

... synergistic interactions between overlapping binding sites for the serum response factor and *ELK-1* proteins ...  
[interact with, serum response factor, *ELK-1*]

---

---

---

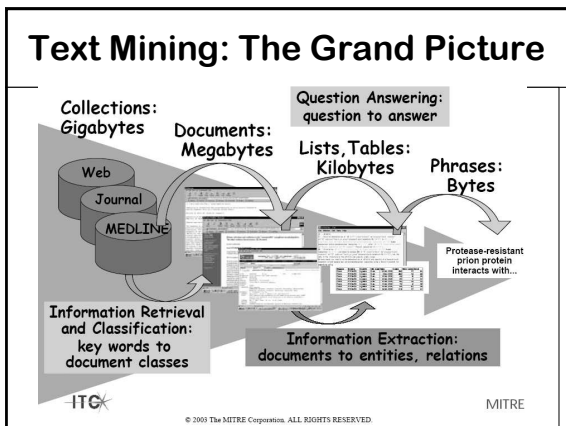
---

---

---

---

---




---

---

---

---

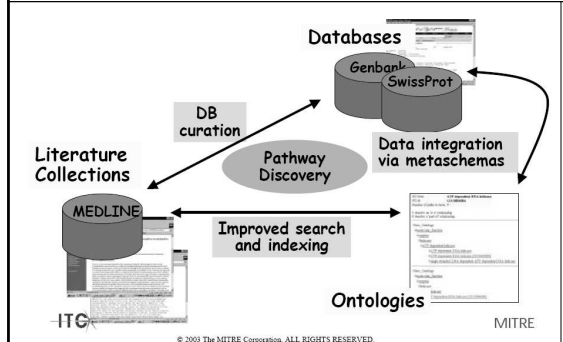
---

---

---

---

## Text Mining as an Integral Part of Biomed Knowledge Management




---

---

---


---

---


---

---

---



## Text Mining



- Data mining on unstructured (NL) or semi-structured (HTML or XML) texts
- Renders 'factoids': text snippets, propositions
- TM is different from information extraction with its emphasis on "new" knowledge
  - not known a priori (even by experts?!)
  - non-redundant
  - some sort of knowledge aggregation (abstraction – textual or visual)
- Knowledge discovery rather than fact retrieval

---

---

---

---

---

---

---

---

## Tutorial Outline

- What is Text Mining?
- Naïve Approach to Text Mining
- Linguistic Approach to Text Mining
- Empirical Approach to Text Mining
- Resources for Empirical NLP
- Summary and Outlook

---

---

---

---

---

---

---

---

## Naïve Approach

- Perform string search of interesting entities and relationships on plain text
- The structure of the query language allows you to search for
  - complete or partial strings
  - string alternatives
  - strings which fulfill simple context conditions
- So, what you are dealing with are, in effect, regular expressions (RegExs)

---

---

---

---

---

---

---

---

## Syntax of Regular Expressions (1/3)

- . masks exactly one single character  
"do." à "dog", "dot", "doe", etc.
- \* zero or more characters may follow  
"do\*" à "do", "dog", "done", "doctor", etc.
- + one or more characters may follow  
"fre+." à "free", "freak", "fresh", "freeze", etc.
- ? zero or exactly the previous character go away  
"ton?e" à "toe" and "tone"

---

---

---

---

---

---

---

---

## Syntax of Regular Expressions (2/3)

- ( ) grouping and disjunction |  
(dog|cat) à "dog" or "cat"
- [ ] any character from the list  
"r[aeiou]t" à "rat", "ret", "rit", "rot", "rut"
- ^ any character not from the list  
t[^aeiou].\*s" à "thanks", "this", "trappings", etc.

---

---

---

---

---

---

---

---

## Syntax of Regular Expressions (3/3)

### - Character Classes -

`\d` any digit [0-9]  
`\D` any non-digit [^0-9]  
  
`\w` any alphanumeric [a-zA-Z0-9\_]  
`\W` any non-alphanumeric [^a-zA-Z0-9\_]  
  
`\s` any space [\t\n\r\f]  
`\S` any non-space [^\t\n\r\f]

---

---

---

---

---

---

---

---

## Demo of Regular Expressions

- It's demo time

---

---

---

---

---

---

---

---

## Shortcomings of RegExs

- Accommodating RegExs to a large variety of NL phenomena makes them increasingly complex
- Though getting more and more complex, RegExs identify not only (some of) the targeted NL expressions but match garbage as well
- More “background knowledge” related to the structure of natural languages might be required

---

---

---

---

---

---

---

---

## Tutorial Outline

- What is Text Mining?
- Naïve Approach to Text Mining
- Linguistic Approach to Text Mining
- Empirical Approach to Text Mining
- Resources for Empirical NLP
- Summary and Outlook

---

---

---

---

---

---

---

---

## Linguistic Approach

- Natural language is described at several layers
  - Lexical layer
    - words
  - Syntactic layer
    - *grammatically coherent* sequences of words in terms of phrases, clauses, and sentences (utterances)
  - Semantic layer
    - meaning of utterances
  - Discourse layer
    - *textually coherent* sequences of utterances as they typically occur in conversations or documents

---

---

---

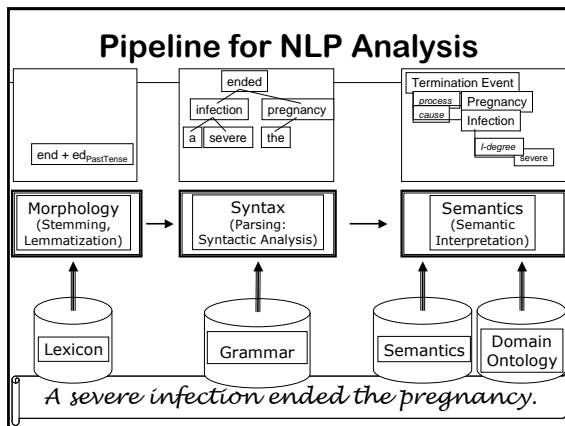
---

---

---

---

---




---

---

---

---

---

---

---

---

### Sample NLP Analysis for Information Extraction

<p><b>Syntactic Analysis &amp; Semantic Tagging</b></p> <p><b>A severe infection ended the pregnancy.</b></p> <p>SUBJ: Phrase: <b>a severe infection</b>; Class: &lt;Disease&gt;</p> <p>VERB: Term: <b>ended</b>; Root: "end" or "terminate"; Mode: active, affirmative</p> <p>OBJ: Phrase: <b>the pregnancy</b> Class: &lt;Process&gt;</p>	<p><b>Information Extraction Rule</b></p> <p>CONCEPT TYPE: <i>Process Termination Event</i></p> <p>CONSTRAINTS:</p> <p>SUBJ: (extract: Cause) Class: &lt;Disease&gt;</p> <p>VERB: Root: "end" or "terminate"; Mode: active</p> <p>OBJ: (extract: PhysProcess) Class: &lt;Process&gt;</p>
<p><b>Extracted Template</b></p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: auto;"> <p>[ <i>Process Termination Event</i> [ Cause : <b>a severe infection</b> ] [ PhysProcess : <b>pregnancy</b> ] ]</p> </div>	

---

---

---

---

---

---

---

---

---

---

### Lexical Layer

- **What is a word, lexically speaking?**
  - "go", "blue", "gene", "take off", "get rid of", "European Union", "on behalf of"
- **Morphology of words**
  - **Inflection**
    - "activate", "activates", "activated"
  - **Derivation**
    - "active", "activate", "activation", "activity"
  - **Compounding**
    - "brain activity", "hyper-activity"

end + ed<sub>PastTense</sub>

---

---

---

---

---

---

---

---

---

---

### Syntactic Layer

- **What is a phrase?**
  - "a severe infection", "the pregnancy"
  - $\neq$  "infection ended the"
- **How do phrases combine to form a clause or a sentence?**

ended

infection

pregnancy

a

severe

the

S

NP

VP

NP

V

NP

Det

Adj

N

V

Det

N

a

severe

infection

ended

the

pregnancy

---

---

---

---

---

---

---

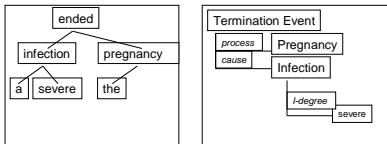
---

---

---

## Semantic Layer

- What is the meaning of a word?
  - “go”, “blue”, “gene”, “take off”, “get rid of”, “European Union”, “on behalf of”
- What is the meaning of a phrase or sentence?



---

---

---

---

---

---

---

---

## Discourse Layer

Thalidomide was found to be highly effective in managing the cutaneous manifestations of leprosy. It was even superior to aspirin in controlling leprosy-associated fever.

---

---

---

---

---

---

---

---

## Discourse Layer

Thalidomide was found to be highly effective in managing the cutaneous manifestations of leprosy. It was even superior to aspirin in controlling leprosy-associated fever.

Otherwise,  
incomplete knowledge gets mined

---

---

---

---

---

---

---

---

**Discourse Layer**

---

Thalidomide was found to be highly effective in managing the cutaneous manifestations of leprosy.

That treatment was even superior to aspirin in controlling leprosy-associated fever.

---

---

---

---

---

---

---

---

**Discourse Layer**

---

Thalidomide was found to be highly effective in managing the cutaneous manifestations of leprosy.

That treatment was even superior to aspirin in controlling leprosy-associated fever.

Otherwise,  
incorrect knowledge gets mined

---

---

---

---

---

---

---

---

**Lexical Background Knowledge**

---

- **Lexicon**
  - Part of speech information
    - "rain" can be a Verb (VB) or a Noun (NN)
  - Inflection patterns (classification)
    - "rain" - "rained" vs. "go" - "went"
  - Syntactic features
    - (in)transitivity of verbs (direct object required or not)
    - Head-modifier specs
      - Head nouns may have a determiner, several adjectives, a quantifier as pre-modifiers; NP or PP as post-modifiers
  - Semantic features
    - Semantic types and relations
      - "aspirin" is a "drug", "drugs" are "man-made artifacts"
    - Selection constraints
      - [Drugs] "cure" [Diseases], [MedPers] "treat" [Patient]

---

---

---

---

---

---

---

---

## Syntactic Background Knowledge

- Grammar (syntax of NL)
  - Set of rules or constraints
    - $S \rightarrow NP VP, NP \rightarrow Det Noun, VP \rightarrow Verb NP$
    - "Number" of HeadNoun determines "Number" of Determiner:
      - "the drugs",  $\neq$  "a drugs"

---

---

---

---

---

---

---

---

## Semantic Background Knowledge

- Semantics (meaning representation of NL)
  - Set of rules or constraints
    - $S \rightarrow NP_1 VP, NP \rightarrow Det Noun, VP \rightarrow Verb NP_2$
    - IF
      - a) the head noun of  $NP_1$  denotes a Disease &
      - b) the head noun of  $NP_2$  denotes a Process &
      - c) the main verb denotes a termination event
    - THEN
      - The following proposition can be instantiated:
        - » TERMINATE ( Disease, Process ) as
        - » TERMINATE ( "infection", "pregnancy" )

---

---

---

---

---

---

---

---

## Sample Morphological Analysis

*„A severe infection ended the pregnancy“*

- Porter Stemmer (lexicon-free):
  - „A“ à a, „severe“ à sever,
  - „infection“ à infect, „ended“ à end,
  - „the“ à the, „pregnancy“ à pregnanc
- Full-form Lexicon:
  - „A“ à a<sub>det</sub>, „severe“ à severe<sub>adj</sub>, „infection“ à infection<sub>noun</sub>, „ended“ à end<sub>verb</sub>,
  - „the“ à the<sub>det</sub>,
  - „pregnancy“ à pregnancy<sub>noun</sub>

---

---

---

---

---

---

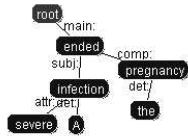
---

---

## Sample Syntactic Analysis

„A severe infection ended the pregnancy.“

- **Constituent Grammar Parser:**  
(S (NP A severe infection)  
(VP ended (NP the pregnancy)) .)
- **Dependency Grammar Parser:**



---

---

---

---

---

---

---

---

## Shortcomings of the “Classical” Linguistic Approach

- Huge amounts of background knowledge req.
  - Lexicons (approx. 100,000 – 150,000 entries)
  - Grammars (>> 15,000 – 20,000 rules)
  - Semantics (>> 15,000 – 20,000 rules)
- As the linguistic and conceptual coverage of classical linguistic systems increases (slowly), it still remains insufficient; systems also reveal ‘spurious’ ambiguity, and, hence, tend to become overly “brittle” and unmaintainable
- More fail-soft behavior is required at the expense of full-depth understanding

---

---

---

---

---

---

---

---

## Tutorial Outline

- What is Text Mining?
- Naïve Approach to Text Mining
- Linguistic Approach to Text Mining
- Empirical Approach to Text Mining
- Resources for Empirical NLP
- Summary and Outlook

---

---

---

---

---

---

---

---

## Empirical Approach

- (Very) large corpora: plain vs. annotated
- Change of mathematics: Statistics rather than algebra and logics
- Linguistic resources: Lexicons, thesauri
- Shallow analysis vs. deep understanding
- Automatic discovery of linguistic regularities rather than manual intuition and calibration — aka NL machine learning
- Large community-wide task-oriented competitions, comparative evaluation rounds

---

---

---

---

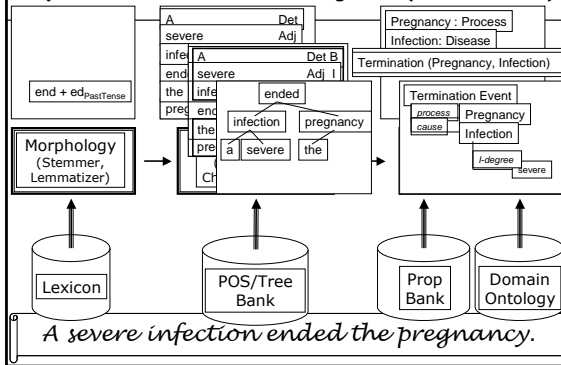
---

---

---

---

## Pipeline for NLP Analysis (revisited)




---

---

---

---

---

---

---

---

## Standard NLP Tools for ...

- Tagging
- Chunking & Partial Parsing
- Named Entity Recognition & Interpretation
- ...

---

---

---

---

---

---

---

---

### Tagging

A severe infection ended the pregnancy .

↓ ↓ ↓ ↓ ↓ ↓ ↓

DET ADJ NOUN VERB DET NOUN ST

---

---

---

---

---

---

---

---

### Penn Treebank Tag Set

Tag	Description	Examples
.	sentence terminator	. ! ?
DT	determiner	all an many such that the them these this
JJ	adjective, numeral	first oiled separable battery-powered
NN	common noun	cabbage thermostat investment
PRP	personal pronoun	herself him it me one oneself theirs they
IN	preposition	among out within behind into next
VB	verb (base form)	ask assess assign begin break bring
VBD	verb (past tense)	asked assessed assigned began broke
WP	WH-pronoun	that what which who whom

In total, 45 tags

---

---

---

---

---

---

---

---

### Transformation Rules for Tagging [Brill, 1995]

- **Initial State:** Based on a number of features, guess the most likely POS tag for a given word:
  - die/DET Frau/NOUN ,/COMMA die/DET singt/VFIN
- **Learn transformation rules to reduce errors:**
  - *Change DET to PREL whenever the preceding word is tagged as COMMA*
- **Apply learned transformation rules:**
  - die/DET Frau/NOUN ,/COMMA die/PREL singt/VFIN

---

---

---

---

---

---

---

---



## Chunking

Arginine methylation of STAT1 modulates IFN induced transcription

---

---

---

---

---

---

---

---

## Chunking

[Arginine methylation of STAT1] modulates [IFN induced transcription]

---

---

---

---

---

---

---

---

## Shallow Parsing

[Arginine methylation of STAT1]<sub>NP</sub> [modulates]<sub>VP</sub> [IFN induced transcription]<sub>NP</sub>

---

---

---

---

---

---

---

---

## Shallow Parsing

[ [Arginine methylation]<sub>NP</sub> [of STAT1]<sub>PP</sub> ]<sub>NP</sub>

[Arginine methylation of STAT1]<sub>NP</sub> [modulates]<sub>VP</sub> [IFN induced transcription]<sub>NP</sub>

---

---

---

---

---

---

---

---

## Shallow Parsing

[ [IFN induced]<sub>AP</sub> [transcription]<sub>N</sub> ]<sub>NP</sub>

[ [Arginine methylation]<sub>NP</sub> [of STAT1]<sub>PP</sub> ]<sub>NP</sub>

[Arginine methylation of STAT1]<sub>NP</sub> [modulates]<sub>VP</sub> [IFN induced transcription]<sub>NP</sub>

---

---

---

---

---

---

---

---

## Deep Parsing

[ [IFN induced]<sub>AP</sub> [transcription]<sub>N</sub> ]<sub>NP</sub>

[ [ [Arginine]<sub>N</sub> [methylation]<sub>N</sub> ]<sub>NP</sub> [of]<sub>P</sub> [STAT1]<sub>N</sub> ]<sub>PP</sub> ]<sub>NP</sub>

[ [Arginine methylation]<sub>NP</sub> [of STAT1]<sub>PP</sub> ]<sub>NP</sub>

[Arginine methylation of STAT1]<sub>NP</sub> [ [modulates]<sub>V</sub> [IFN induced transcription]<sub>NP</sub> ]<sub>VP</sub>

---

---

---

---

---

---

---

---



## BIO Format for Base NPs

a	DT	I
mechanism	NN	I
that	WDT	B
increases	VBZ	O
NF-kappa	NN	I
B/I	NN	I
kappa	NN	I
B	NN	I
dissociation	NN	I
without	IN	O
affecting	VBG	O
the	DT	I
NF-kappa	NN	I
B	NN	I
translocation	NN	I
step	NN	I

---

---

---

---

---

---

---

---

---

---

---

---

## A Simple Chunking Technique

- Simple chunkers usually ignore lexical content
  - Only need to look at part-of-speech tags
- Basic steps in chunking
  - Chunking / Unchunking
  - Chinking
  - Merging / Splitting

---

---

---

---

---

---

---

---

---

---

---

---

## Chunking

- Define a regular expression that matches the sequences of tags in a chunk
  - <DT>? <JJ>\* <NN.>
- Chunk all matching subsequences
  - A/DT red/JJ car/NN ran/VBD on/IN t&e/DT street/NN
  - [A/DT red/JJ car/NN] ran/VBD  
on/IN [t&e/DT street/NN]
- If matching subsequences overlap, the first one gets priority
- Unchunking is the opposite of chunking

---

---

---

---

---

---

---

---

---

---

---

---

## Chinking

- A chink is a subsequence of the text that is not a chunk
- Define a regular expression that matches the sequences of tags in a chink
  - (<VB.??> | <IN> )+
- Chunk anything that is *not* a matching subsequence
  - A/DT red/JJ car/NN ran/VBD on/IN the/DT street/NN
  - [A/DT red/JJ car/NN]  
    ran/VBD on/IN the/DT street/NN  
    chink

---

---

---

---

---

---

---

---

## Merging

- Combine adjacent chunks into a single chunk
- Define a regular expression that matches the sequences of tags on both sides of the point to be merged
  - Merge a chunk ending in "JJ" with a chunk starting with "NN", i.e. left: <JJ>, right: <NN.>
- Chunk all matching subsequences
  - [A/DT red/JJ ] [ car/NN] ran/VBD  
    on/IN the/DT street/NN
  - [A/DT red/JJ car/NN] ran/VBD  
    on/IN the/DT street/NN
- Splitting is the opposite of merging

---

---

---

---

---

---

---

---

## Approaches to Chunking

- Hand-coded Grammars (RegEx, FSA)
  - [Abney, 1991]
- Transformation Rule Learning: *TBL*
  - [Ramshaw & Marcus, 1995]
- Hidden Markov Models
  - [Molina & Pla, 2002]
- Support Vector Machines: *YamCha*
  - [Kudo & Matsumoto, 2001]

---

---

---

---

---

---

---

---

## Named Entity Recognition & Interpretation

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> leprosy </DISEASE> (<DISEASE> erythema nodosum leprosum </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<DISEASE> acetylsalicylic acid </DISEASE>) in controlling <DISEASE> leprosy-associated fever </DISEASE>

---

---

---

---

---

---

---

---

## Named Entity Recognition & Interpretation

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> leprosy </DISEASE> (<DISEASE> erythema nodosum leprosum </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<DISEASE> acetylsalicylic acid </DISEASE>) in controlling <DISEASE> leprosy-associated fever </DISEASE>

---

---

---

---

---

---

---

---

## What are Named Entities?

- Names of persons
    - Dr. Jonathan Peafo, Professor Johns
  - Names of companies
    - Sony, Pa
  - Names of locations
    - London, #0 105
  - Dates
    - 1995
  - Addresses
    - udo.hah
    - il-jenä.de
  - Names of proteins or genes or diseases,
    - chloramphenicol acetyltransferase, NF-kappa B, SARS
  - Measure expressions
    - 420 kp, 21 l/m<sup>2</sup>, 37%, 900€
- named entities are intentionally excluded from the lexicon

---

---

---

---

---

---

---

---

## Two Types of NER Methods

### Human Knowledge Engineering

- rule based
- developed by experienced language engineers
- based on human intuition
- requires only small amount of training data
- development could be very time consuming
- some changes may be hard to accommodate

### (Supervised) Machine Learning Systems

- use statistics or other machine learning technique
- developers do (almost) not need linguistic expertise
- fully automatic
- requires large amounts of annotated training data
- annotators are cheap (but you get what you pay for!)
- some changes may require re-annotation of the entire training corpus

---

---

---

---

---

---

---

---

## Naïve NER Method: List Look-up

- Recognize entities stored in given lists (*gazetteers*, e.g., online phone directories, yellow pages)
- Advantages:
  - Simple, fast, language independent, easy to retarget (just create lists)
- Disadvantages:
  - impossible to enumerate all names and name variants, collection and maintenance of lists

---

---

---

---

---

---

---

---

## NER by Pattern Recognition

- Names often have internal structure - these components can be either stored or guessed, e.g., for location we have RegEx-style constraints such as:

Capitalized Word + {City, Forest, Center, River}

which yields: *Sherwood Forest, Manchester City*

Capitalized Word + {Street, Boulevard, Avenue, Road}

which yields: *Portobello Street, Washington Avenue*

---

---

---

---

---

---

---

---

## NER by Expressive Rules

- Context-sensitive rules of the kind:

$A \rightarrow B \setminus C / D$

- A is a set of attribute-value expressions and optional score, the attributes refer to elements of the input token feature vector
- B, C, D are sequences of attribute-value pairs and regular expressions; variables are also supported
- B and D are left and right context, respectively, and can be empty

Example: `[syn=NP, sem=ORG] (0.9) →  
 \ [norm="university"], [token="of"],  
 [sem=REGION|COUNTRY|CITY] / ;`

---

---

---

---

---

---

---

---

## Sample NER Rule

Rule for the mark up of person names when the first name is not present or known from the gazetteers: e.g., "Mr. J. Cass",

```
[ SYN=PROP, SEM=PER,  
  FIRST=_F, INITIALS=_I, MIDDLE=_M, LAST=_L ]  
#_F, _I, _M, _L are variables, transfer info from RHS  
→ [SEM=TITLE_MIL | TITLE_FEMALE | TITLE_MALE]  
  \ [SYN=PROP, ORTH=CAP[A, TOKEN=_F]?,  
    [SYN=NAME, ORTH=INIT[O, TOKEN=_I]?,  
    [SYN=NAME, ORTH=INIT[O, TOKEN=_M]?,  
    [SYN=PROP, ORTH=CAP[A]O, TOKEN=_L, SOURCE≠RULE]  
  #proper name, not recognised by a rule  
  / ;
```

---

---

---

---

---

---

---

---

## NER by Machine Learning

- NE task is frequently broken down in two parts:
  - Recognizing the entity boundaries
  - Classifying the entities in the NE categories
- Features are at least as important as the choice of the ML method
  - Simple pattern matching of orthographic features: capitalization, punctuation marks, numerical symbols
  - Windows for lexical features (e.g., "Mr." for persons)
  - Affix features ("-ase" for proteins, "-ectomy" for medical procedures, etc.)
  - POS info (and chunks)
- Major Approaches
  - Maximum Entropy [Chieu & Ng, 2002]
  - Hidden Markov Models [Bikel et al., 1999]
  - Support Vector Machines [Takeuchi & Collier, 2002]

---

---

---

---

---

---

---

---





## Bad News from the BioMed Field

Analysis of DNA methylation revealed rice adult plants resistant to bacterial blight based on methylation-sensitive AFLP (MSAP) analysis.

Sha AH, ~~Wu~~ XH, Huang JB, Zhang DP.

National Key Laboratory of Crop Genetic Improvement, National Center of Crop Molecular Breeding, Huazhong Agricultural University, Wuhan, 430070, China, [ashang@caas.hzau.edu.cn](mailto:ashang@caas.hzau.edu.cn)

DNA methylation is known to play an important role in the regulation of gene expression in eukaryotes. The rice cultivar Wase Aikoku 3 becomes resistant to the blight pathogen *Xanthomonas oryzae* pv. *oryzae* at the adult stage. Using methylation-sensitive amplified polymorphism (MSAP) analysis, we compared patterns of cytosine methylation in seedlings and adult plants of the rice cultivar Wase Aikoku 3 that had been inoculated with the pathogen *Xanthomonas oryzae* pv. *oryzae*, subjected to mock inoculation or not inoculated. In all, 888 DNA fragments, each representing a recognition site cleaved by either or both of two isoschizomers, were amplified using 60 pairs of selective

<input checked="" type="checkbox"/>	Address
<input checked="" type="checkbox"/>	Date
<input type="checkbox"/>	FirstPerson
<input type="checkbox"/>	Location
<input type="checkbox"/>	Lookup
<input checked="" type="checkbox"/>	Organization
<input checked="" type="checkbox"/>	Person
<input type="checkbox"/>	SpaceToken
<input type="checkbox"/>	Token
<input type="checkbox"/>	Original markups

---

---

---

---

---

---

---

---

---

---

## PASTA – NER Examples

Staphylococcus aureus enterotoxin A (SEA) belongs to a subgroup of the staphylococcal superantigens that utilizes Zn<sup>2+</sup> in the high affinity interaction with MHC class II molecules. A high affinity metal binding site was described previously in SEA cocrystallized with Cd<sup>2+</sup> in which the metal ion was octahedrally coordinated, involving the N-terminal serine.

---

---

---

---

---

---

---

---

---

---

## PASTA – NER Examples

Staphylococcus aureus enterotoxin A (SEA) belongs to a subgroup of the staphylococcal superantigens that utilizes Zn<sup>2+</sup> in the high affinity **SPECIES** MHC class II molecules. A high affinity metal binding site was described previously in SEA cocrystallized with Cd<sup>2+</sup> in which the metal ion was octahedrally coordinated, involving the N-terminal serine.

---

---

---

---

---

---

---

---

---

---

## PASTA – NER Examples

Staphylococcus aureus enterotoxin A ( SEA ) belongs to a subgroup of the staphylococcal superantigens that utilizes Zn<sup>2+</sup> in the high affinity interaction with MHC class II molecules. A high affinity metal binding site was described previously in SEA cocry **PROTEIN** in which the metal ion was coordinated, involving the N-terminal serine .

---

---

---

---

---

---

---

---

## Document Processing in GATE/PASTA

- Tokenizer
- Lemmatizer
- NER: Gazetteer Lookup
- Sentence Splitter
- POS-Tagger
- Semantic Tagger
- Name Matcher (Named Entity Recognition)
- Parser (AVM: Attribute Value Matrix)
- Extraction Rules




---

---

---

---

---

---

---

---

## Information Extraction in PASTA

*Results: We have determined the crystal structure of a triacetylgllycerol lipase from Pseudomonas cepacia (Pet) in the absence of a bound inhibitor using X-ray crystallography. The structure shows the lipase to contain an alpha/beta-hydrolase fold and a catalytic triad comprising of residues Ser221, His386 and Asp264. The enzyme shares several structural features with homologous lipases from Pseudomonas glumae (PgL) and Chromobacterium viscosum (CvL), including a calcium-binding site. The present structure of Pet reveals a highly open conformation with a solvent-accessible active site. This is in contrast to the structures of PgL and Pet in which the active site is buried under a closed or partially opened lid, respectively.*

```

<RESIDUE-str-1997-5-2-1>:
RESIDUE_TYPE: SERINE
RESIDUE_NO: "87"
IN_PROTEIN: <PROTEIN-str-1997-5-2-1>
SITE/FUNCTION: "active site", "catalytic",
"interfacial activation", "calcium-binding site"
SECOND_STRUCT: alpha-helix
REGION: 'lid'
ARTICLE: <ARTICLE-str-1997-5-2-1>

<PROTEIN-str-1997-5-2-1>:
NAME: "Triacylglycerol lipase"
SCOP_CLASS: "Lipase"
PDB_CODE: 1LGY
IN_SPECIES: <SPECIES-str-1997-5-2-1>

<SPECIES-str-1997-5-2-1>:
NAME: "Pseudomonas cepacia"
NAME_TYPE: SCIENTIFIC
    
```

---

---

---

---

---

---

---

---

## Challenges for NER

- Identify which mentions in the text refer to which entities
  - *Tony Blair, Mr. Blair, he, the prime minister*
- Semantic tagging of entities
  - Expanding the set of entities to be recognized – e.g., substances (food, drugs, genes, proteins, products)
  - Finer-grained entity hierarchies
    - Organizations (governmental, commercial, educational, etc.),
    - Locations (regions, countries, cities, rivers, lakes, seas, etc.)

---

---

---

---

---

---

---

---

## Hierarchy-Building Relations in Biomedical Ontologies

- Taxonomy: *Is-A* ( $C_1, C_2$ )



---

---

---

---

---

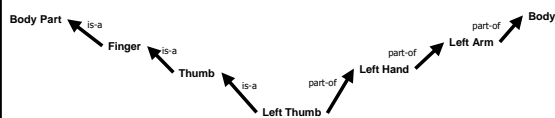
---

---

---

## Hierarchy-Building Relations in Biomedical Ontologies

- Taxonomy: *Is-A* ( $C_1, C_2$ )
- Partonomy: *Part-of* ( $C_1, C_2$ )



---

---

---

---

---

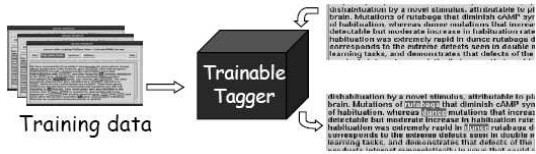
---

---

---



## Infrastructure Requirements



- **Manual Creation of Corpora**
  - Training Coders in Description Languages
  - Test of Coder Reliability
- **Benefit:**
  - Solid Foundation for Supervised Learning

---

---

---

---

---

---

---

---

## General Language Corpora w./ Syntactic, Semantic Annotations

- **Penn Treebank (U Penn)** [Marcus et al., 1993]
  - language: English
  - size: 1,200,000 (POS- & tree-annotated) tokens
  - text genre: mostly newspaper articles (*Wall Street Journal*)
  - tag set of 45 tags; phrase structure grammar
- **Penn PropBank (U Penn)** [Palmer et al., 2005]
  - size: 300,000 (annotated) tokens
  - text genre: financial newspaper articles (*Wall Street Journal*)
  - proposition format: pred(arg0, arg1, arg2, ...)

---

---

---

---

---

---

---

---

## Sublanguage Corpora w./ Syntactic Annotations

- **GENIA (U Tokyo)** [Ohta et al., 2002]
  - language: English
  - size: 2.000 abstracts (18,500 sent.s, 491,000 tokens)
    - selected from a MeSH term search of "Human", "Blood Cells" and "Transcription Factors"
  - text genre: biology articles (*Medline* bibliographic db)
  - tag set: Penn TreeBank (PTB)
  - beta version of PTB-style treebank (200 abstracts)

---

---

---

---

---

---

---

---

## Sublanguage Corpora w./ Named Entity Recognition

- GENIA (U Tokyo)
  - language: English
  - content: subset of substances (peptides, amino acids, DNA), biological locations (organisms, tissues) involved in reaction of proteins (GENIA ontology) — 100,000 bio annotations
- BioCreative (MITRE, USA)
  - size: 15,000 sentences; ~ 397,000 tokens
  - content: gene and protein names

---

---

---

---

---

---

---

---

## General Lexicons

- WordNet (Princeton U)
  - canonical word forms (nouns, verbs, adjectives etc.)
  - lexical semantics: taxonomies & partonomies of synonym classes (synsets), definitions (glosses)
  - language: English (120,000 entries)
  - content: general language
- FrameNet (Berkeley U)
  - more semantically oriented (frame semantics)

---

---

---

---

---

---

---

---

## Sublanguage Lexical Resources

- Unified Medical Language System, UMLS (NLM)
  - Umbrella system made up of approximately 100 terminologies, including *Gene Ontology* (GO)
  - Basic and variant word forms, and (quite complex) noun phrases
  - Lexical semantics: thesaurus relations for taxonomies, partonomies, also other light-weight semantics
  - Language: English
  - Size: 1,000,000 terms/concepts, 11,000,000 relations
  - Content: (almost) the whole (bio)medical domain
  - (English) Specialist Lexicon uses conceptual grounding of UMLS for NLP applications

---

---

---

---

---

---

---

---

## Sublanguage Lexical Resources

- **Gene Ontology (GO Consortium)**
  - Terms with lexical definitions and semantic relations (taxonomy, partonomy)
  - Three basic relational ontologies: *cellular component*, *molecular function*, *biological process*
  - Size: 17,600 terms (93% w./ definitions)
  - GO  $\xi$  GONG (DAML+OIL-based description logics version)

---

---

---

---

---

---

---

---

## Domain Ontologies

- **CoMeT (U Freiburg / U Jena)**
  - description logics (classifier, realizer)
  - LOOM implementation ( $\xi$  OWL)
  - language-independent
  - exported from the whole anatomy and pathology section of the UMLS
  - size: 240,000 concepts and relations
- **\*NCI\* (US National Cancer Institute)**
  - description logics (classifier)
  - OWL implementation
  - language-independent
  - oncology
  - size: 26,000 concepts, 71,000 terms

---

---

---

---

---

---

---

---

## Demo of GENIA

### Example:

*„Preincubation of cells with 1,25-(OH)<sub>2</sub>D<sub>3</sub> augmented IL-1 beta mRNA levels only in U-937 and HL-60 cells.“*

---

---

---

---

---

---

---

---

## POS Annotation in GENIA

Preincubation/NN of/IN cells/NNS  
with/IN 1,25-(OH)2D3/NN  
augmented/VBD IL-1/NN beta/NN  
mRNA/NN levels/NNS only/RB in/IN U-  
937/NN and/CC HL-60/NN cells/NNS ./.

---

---

---

---

---

---

---

---

## Named Entity Annotation in GENIA

```
- <sentence>
  Preincubation of cells with
  <cons lex="1,25-(OH)2D3" sem="G#lipid">1,25-(OH)2D3</cons>
  augmented
  - <cons lex="IL-1_beta_mRNA_level" sem="G#other_name">
  - <cons lex="IL-1_beta_mRNA" sem="G#RNA_molecule">
    <cons lex="IL-1_beta" sem="G#protein_molecule">IL-1 beta</cons>
  mRNA
  </cons>
  levels
  </cons>
  only in
  <cons lex="U-937" sem="G#cell_line">U-937</cons>
  and
  <cons lex="HL-60" sem="G#cell_line">HL-60</cons>
  cells.
</sentence>
```

---

---

---

---

---

---

---

---

## Syntactic Annotation in GENIA

```
- <S>
- <NP-SBJ>
  <NP> Preincubation/NN</NP>
- <PP>
  of/IN
  <NP> cells/NNS</NP>
</PP>
- <PP>
  with/IN
  <NP> 1,25-(OH)2D3/NN</NP>
</PP>
</NP-SBJ>
- <VP>
  augmented/VBD
  <NP> IL-1/NN beta/NN mRNA/NN levels/NNS</NP>
- <PP>
  only/RB in/IN
- <NP>
  - <NP SYN="COORD">
    <NP> U-937/NN</NP>
  and/CC
  <NP> HL-60/NN</NP>
</NP>
  cells/NNS
</NP>
</PP>
</VP>
</S>
```

---

---

---

---

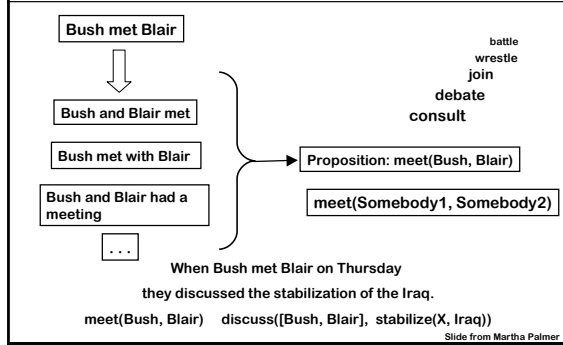
---

---

---

---

## Example for Propositions (PPB)




---

---

---

---

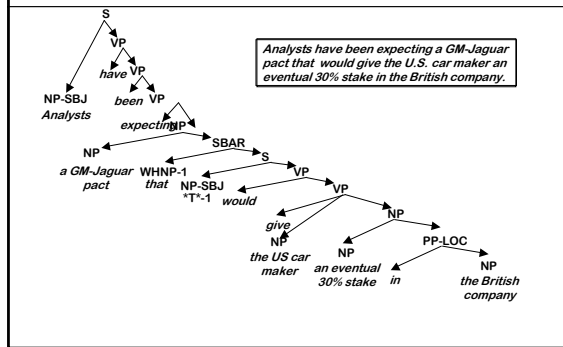
---

---

---

---

## PTB Treebank Sentence




---

---

---

---

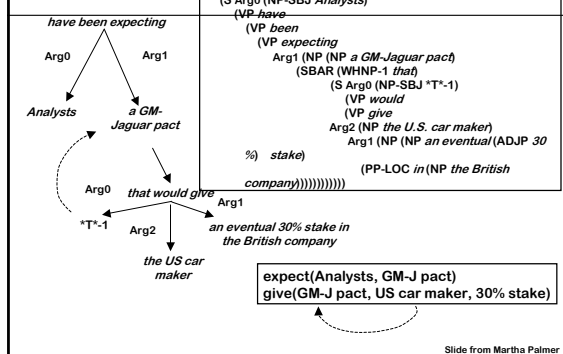
---

---

---

---

## Sentence in PropBank-Format




---

---

---

---

---

---

---

---

## Different Meanings of a Verb

<i>Mary called John an idiot.</i> (LABEL)	<i>Mary called John a cab.</i> (SUMMON) <sup>5</sup>
Arg0: Mary	Arg0: Mary
Rel: called	Rel: called
Arg1: John (item being labeled)	Arg2: John (benefactive)
Arg3-PRD: an idiot (attribute)	Arg1: a cab (thing summoned)

---

---

---

---


---

---

---

---

## Where Are We Now?

- State of annotated biomedical corpora
  - good coverage of tagging annotations
  - little coverage of chunking/parsing annotations (mind their quality)
  - reasonable coverage of named entities (ontology!)
  - no coverage of structured proposition annotations
  - no coverage of discourse annotations (zoning)
- Quality control is a (unsolved) key is 
  - single coders (!)
- Size matters
- Community-wide efforts should be started
  - Working groups composed of biologists, computer scientists, computational linguists, ontologists

---

---

---

---

---

---

---

---

## How Good Are We?

 Granada, Mar 28-31, 2004

- Critical Assessment of Information Extraction
  - 1<sup>st</sup> critical assessment of text mining in biology
  - Run by CNB, Madrid (Valencia, Blaschke) + MITRE (Hirschman, Yeh, Colosimo, Morgan)
  - Supported by NSF (MITRE) and EMBO (CNB)
  - 27 groups participated from 10 countries
- Focus on tasks relevant to biologists:
  - Task 1: Gene name extraction and normalization
    - Identification of gene names in PubMed abstracts & normalization to unique gene IDs
  - Task 2: Functional annotation
    - Identification of textual evidence and GO codes, for a given protein, in full text articles

© 2004 The MITRE Corporation. ALL RIGHTS RESERVED.

MITRE

---

---

---

---

---

---

---

---

## How Good Are We?

### Task 1A Overview: Gene Name Finding

- Identify all mentions of genes, proteins, etc in sentences from PubMed abstracts
  - Data provided by Lorrie Tanabe & John Wilbur, NCBI/NLM; task run by Alex Yeh, MITRE

Mutation of TTF-1-binding sites (TBE) 1, 3, and 4 in combination markedly decreased transcriptional activity of SP-A promoter-chloramphenicol acetyltransferase constructs containing SP-A gene sequences from -256 to +45.

- Note that mentions can be complex and long!
- Gold standard created manually by biologists
- Building block for more complex tasks

---

---

---

---

---

---

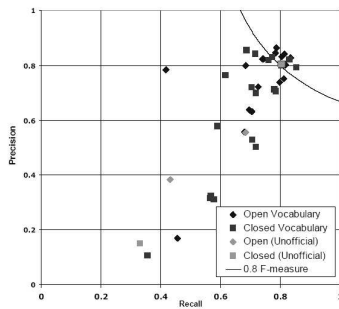
---

---

---

---

### Task 1A Results: 4 Groups > 80% F-score



14+1 groups submitted 44 runs

High scores:  
Recall 0.85  
Precision 0.86  
F-score 0.83

F-score =  $2 * R * P / (R + P)$   
Recall =  $TP / (TP + FN)$   
Precision =  $TP / (TP + FP)$

© 2004 The MITRE Corporation. ALL RIGHTS RESERVED.

MITRE

---

---

---

---

---

---

---

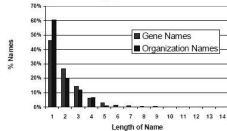
---

---

---

### Task 1A Lessons Learned: Longer Names Are Harder

Results comparable to previous published bio results  
But still lag behind 0.93 F-score for newswire - why?



Gene names longer than organization names:

Gene: 2.1wds avg  
Org: 1.7 wds avg

- Organization name (MUC)
  - over all words: F-score 0.93
  - estimated single word: F-score ~ 0.96
- Gene name (BioCreAtIvE)
  - over all words: F-score 0.83
  - estimated single-word: F-score ~ 0.92

© 2004 The MITRE Corporation. ALL RIGHTS RESERVED.

MITRE

---

---

---

---

---

---

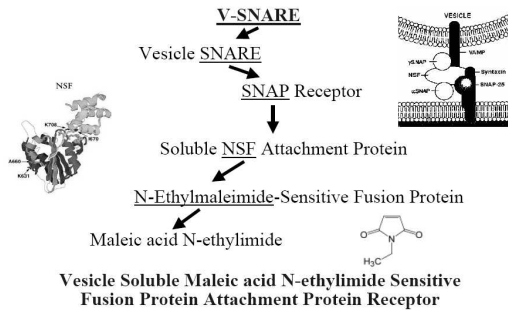
---

---

---

---

## Why Bio Terminology is so Hard



MITRE

Copyright 2005, MITRE Corporation

---

---

---

---

---

---

---

---

---

---

## Quantitative Results for Relation Mining

Source	Relation	Entity	DB	Prec	Recall
Craven '99	location	protein	Yeast	92%	21%
Rindfleisch '99	binding	UMLS	MEDLINE	79%	72%
Proux '00	interact	gene	Flybase	81%	44%
Pustejovsky '02	inhibit	gene	MEDLINE	90%	57%
Friedman '01	pathway	many	Articles	96%	63%

○ Tasks differ greatly, e.g., finding pathway information (Friedman '01) is more general than finding just binding relations (Rindfleisch '99)

○ We cannot compare these different approaches unless we define a common "challenge evaluation"

---

---

---

---

---

---

---

---

---

---

## Tutorial Outline

- What is Text Mining?
- Naïve Approach to Text Mining
- Linguistic Approach to Text Mining
- Empirical Approach to Text Mining
- Resources for Empirical NLP
- Summary and Outlook

---

---

---

---

---

---

---

---

---

---

## Summary

- Biomedical text mining is a very active field
- Danger of too many people using too weak methods (RegEx)
- Empirical NLP provides off-the-shelf infrastructure for core NLP processes (morphological, syntactic, & semantic)
- Bio named entities are extremely hard to identify and to interpret
- Lacking biomed resources for semantics
- Dramatic ontology gap (for bio, in particular)
- Almost no support for genuine TM tasks such as redundancy minimization, flagging for newsworthiness, abstraction layers, etc.

---

---

---

---

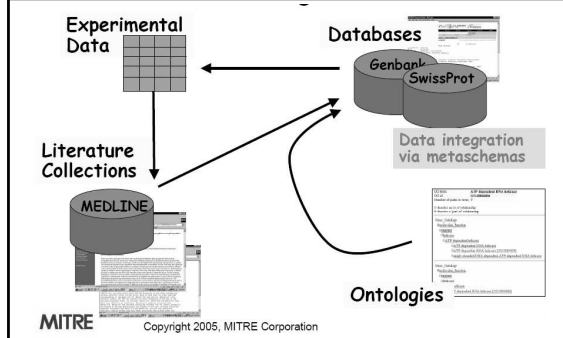
---

---

---

---

## Text Mining is just a Part of it ...



---

---

---

---

---

---

---

---

## Yet Another Conclusion

*"About a quarter of late stage failures we surveyed could have been eliminated two years earlier by making all internal information in the form of documents more widely available."*

Top 5 Pharmaceutical Senior VP  
MBC Informatics Committee, July 31, 2003

---

---

---

---

---

---

---

---

semantic interoperability and data mining in biomedicine  
SEMANTIC MINING

Information Society Technologies

# Text Mining

Tutorial at the 2005 NoE "Semantic Mining" Summer School  
Tihany, Hungary, July 2, 2005

Udo Hahn & Michael Poprat

FRIEDRICH-SCHILLER-UNIVERSITÄT  
JENA  
www.coling.uni-jena.de

ena  
niversity  
anguage  
and  
nformation  
engineering

---

---

---

---

---

---

---

---

## Recommended Readings

- **Textbooks on Natural Language Processing**
  - D. Jurafsky & J.A. Martin (2000), *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.
  - C.D. Manning & H. Schütze (1999), *Foundations of Statistical Natural Language Processing*. MIT Press.
- **Text Mining (for Biology)**
  - P. Jackson & I. Mouliner (2002), *Natural Language Processing for Online Applications, Text Retrieval, Extraction and Classification*. John Benjamins.
  - B.J. Stapley & S. Ananiadou (forthcoming), *Text Mining for Biology*. Artech House Books.
  - H. Shatkay & R. Feldman (2003). Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821-855.
  - L. Hirschman, J.C. Park, J.-i. Tsujii, L. Wong & C. Wu (2002), Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553-1561.

---

---

---

---

---

---

---

---

## Advanced Readings

- **Tagging**
  - E. Brill (1995), Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543-565.
  - T. Brants (2000), TnT - a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pp.224-231.
  - A. Ratnaparkhi (1996), A maximum entropy part-of-speech tagger. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*, pp.133-141.
- **Chunking**
  - S. Abney (1991), Parsing by chunks. In R. Berwick, S. Abney & T. Carol (Eds.), *Principle-Based Parsing*. Kluwer, pp.257-278.
  - R. Ramshaw & M.P. Marcus (1995), Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pp.82-94.
  - A. Molina & F. Pla (2002), Shallow parsing using specialized HMMs. *Journal of Machine Learning Research*, 2(4):595-613.
  - T. Kudo & Y. Matsumoto (2001), Chunking with Support Vector Machines. In *NAACL '01 - Language Technologies 2001, Proceedings of the Second Meeting of the North American Chapter of the ACL*, pp.192-199.

---

---

---

---

---

---

---

---

## Advanced Readings

- **Named Entity Recognition**
  - D. Bikel, R. Schwartz & R. Weischedel (1999), An algorithm that learns what's in a name. *Machine Learning*, 34(1/3):211-231.
  - H. Chieu & H.T. Ng (2002), Named entity recognition: A maximum entropy approach using global information. In *Proceedings of the 19th International Conference on Computational Linguistics - COLING 2002*, pp.190-196.
  - K. Takeuchi & N. Collier (2002), Use of Support Vector Machines in extended named entity recognition. In *Proceedings of the Sixth Conference on Natural Language Learning - CoNLL 2002*, pp.119-125.
- **Textbooks on Machine Learning**
  - T. Mitchell (1997), *Machine Learning*. McGraw Hill.
  - N. Christianini & J. Shawe-Taylor (2001), *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

---

---

---

---

---

---

---

---

## Advanced Readings

- **(Bio) NLP Resources**
  - M.P. Marcus, B. Santorini & M.A. Marcinkiewicz (1993), Building a large annotated corpus of English: The PENN TREEBANK. *Computational Linguistics*, 19(2):313-330.
  - M. Palmer, D. Gildea & P. Kingsbury (2005), The PROPOSITION BANK: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71-105.
  - T. Ohta, Y. Tateisi & J.-D. Kim (2002), The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *HLT 2002 - Proceedings of the Second International Human Language Technology Conference*, pp.82-86.
- **Biomedical Information Extraction Systems**
  - A. Rzhetsky et al. (2004), GENEWAYS: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43-53.
  - C. Friedman et al. (2001), GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Supplement 1):S74-S82.
  - R. Gaizauskas et al. (2003), Protein structures and information extraction from biological texts: The PASTA system. *Bioinformatics*, 19(1):135-143.

---

---

---

---

---

---

---

---