

# **Linking Text with Knowledge**

## **Challenges in Bio-Text Processing**

**Junichi TSUJII**

**School of Computer Science  
National Centre for Text Mining**

**University of Manchester, UK**

**Department of Computer Science  
School of Information Science and  
Technology**

**University of Tokyo, JAPAN**

# NLP and TM

## Natural Language Processing

Language as a complex system linking surface strings of characters with their meanings  
Text and words as structured objects

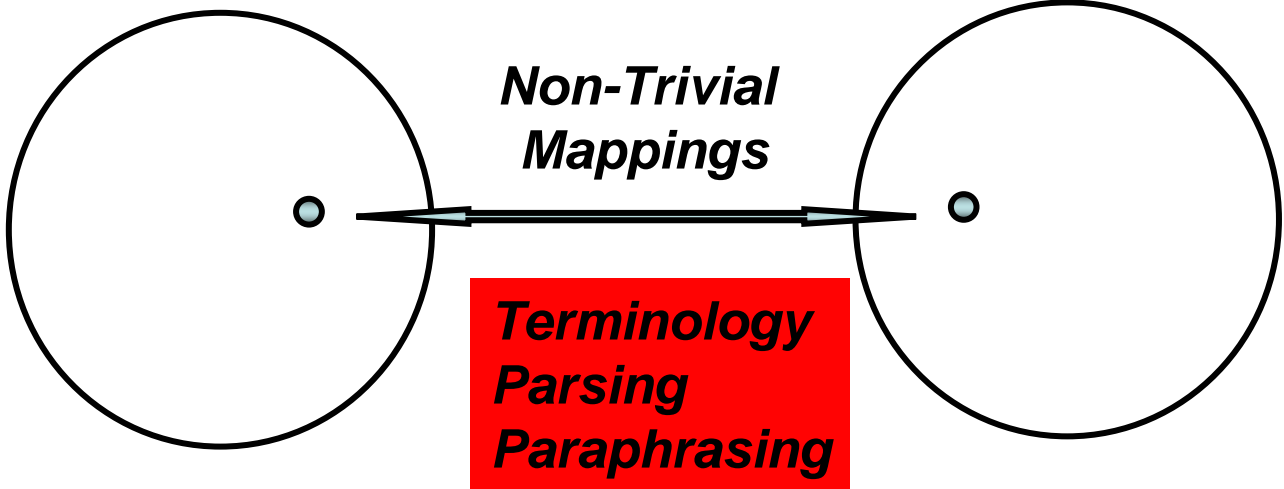
## Text Mining

Text as a bag of words  
Words as surface strings



NLP-based TM

**From surface diversities and ambiguities  
to  
conceptual invariants**



**Language Domain**

**Knowledge Domain**

**Linguistic expressions**

**Concepts and Relationships  
among Them**

**Motivated  
Independently of language**



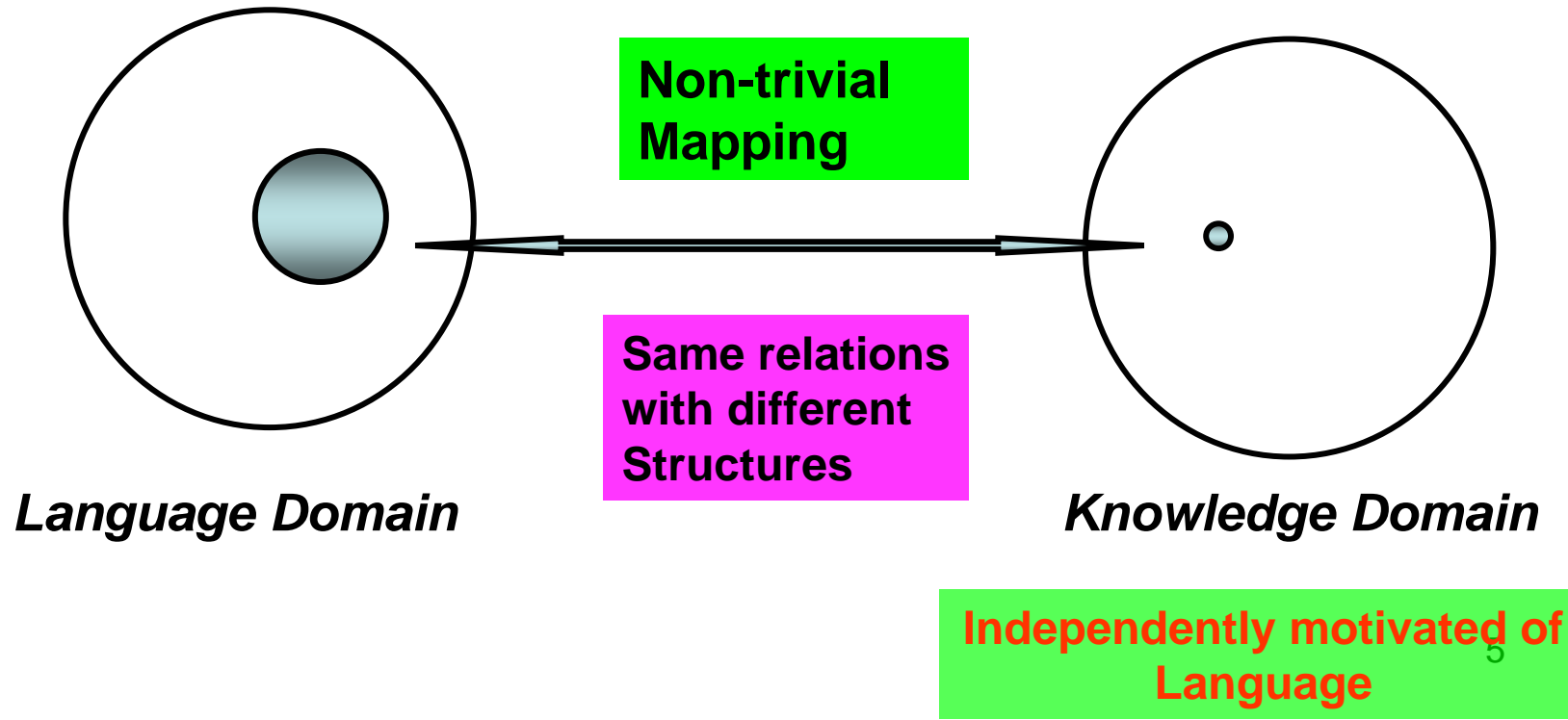
# Example

[A] protein activates [B] (Pathway extraction)

Transcription initiation by the sigma(54)-RNA polymerase holoenzyme requires an **enhancer-binding protein** that is thought to contact sigma(54) to **activate** transcription. **Full-strength** **Streptococcus pneumoniae** **phosphorylated** **enhancer-binding protein** could activate transcription, but fails to associate with **oskR** RNA-protein activation surfaces of P.HO5 gene.

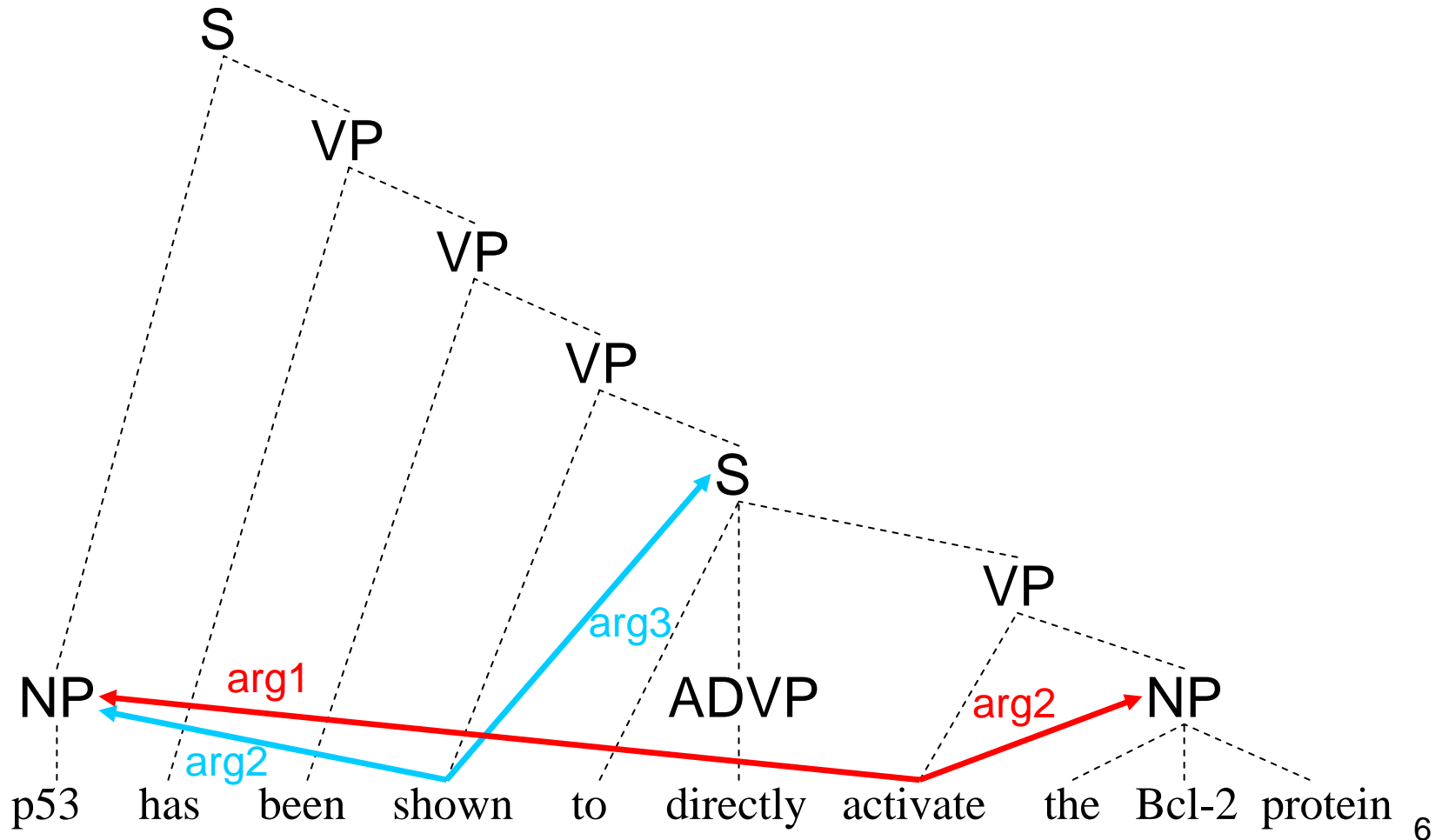
[sentence] > ([arg1\_activate] > [protein])

Retrieval using Regional Algebra



# Predicate-argument structure

Parser based on Probabilistic HPSG (Enju)



Semantic Search **Keyword Search** GCL Search

subject	verb	object
<input type="text" value="p53"/>	<input type="text" value="activate"/>	<input type="text"/>

# Semantic Retrieval System Using Deep Syntax MEDIE

Results 1-50 for **p53 activate** [Show next](#) [Show query](#)

- [PMID: 15446548](#) [XML](#)  
The molecules activated by p53 induce apoptosis, cell cycle arrest, and DNA repair to conserve genome.
- [PMID: 15273740](#) [XML](#)  
In this report, we demonstrated that human AMID gene promoter was activated by p53 in reporter gene assays.
- [PMID: 15020844](#) [XML](#)  
Recently, p53 has been shown to directly activate the pro-apoptotic Bcl-2 protein.
- [PMID: 15105421](#) [XML](#)  
Electrophoretic mobility shift assays reveal that both transcription factors are capable of binding to putative consensus sites, and luciferase reporter assays reveal that E2F1 and p53 can activate transcription from the SIVA promoter.
- [PMID: 15247038](#) [XML](#)  
Although the role of the nuclear factor-kappa B (NF-kappa B) signaling cascade is crucial in ICAM-1 activation, we have shown that p53 directly activates the expression of ICAM-1 in an NF-kappa B-independent manner.
- [PMID: 15021899](#) [XML](#)  
Because the MDM2 gene is transcriptionally activated by p53, it forms part of an autoregulatory feedback loop that directly links the transcriptional activity of p53 with its degradation.
- [PMID: 15064739](#) [XML](#)

Passive

Passive and Infinitival Clause



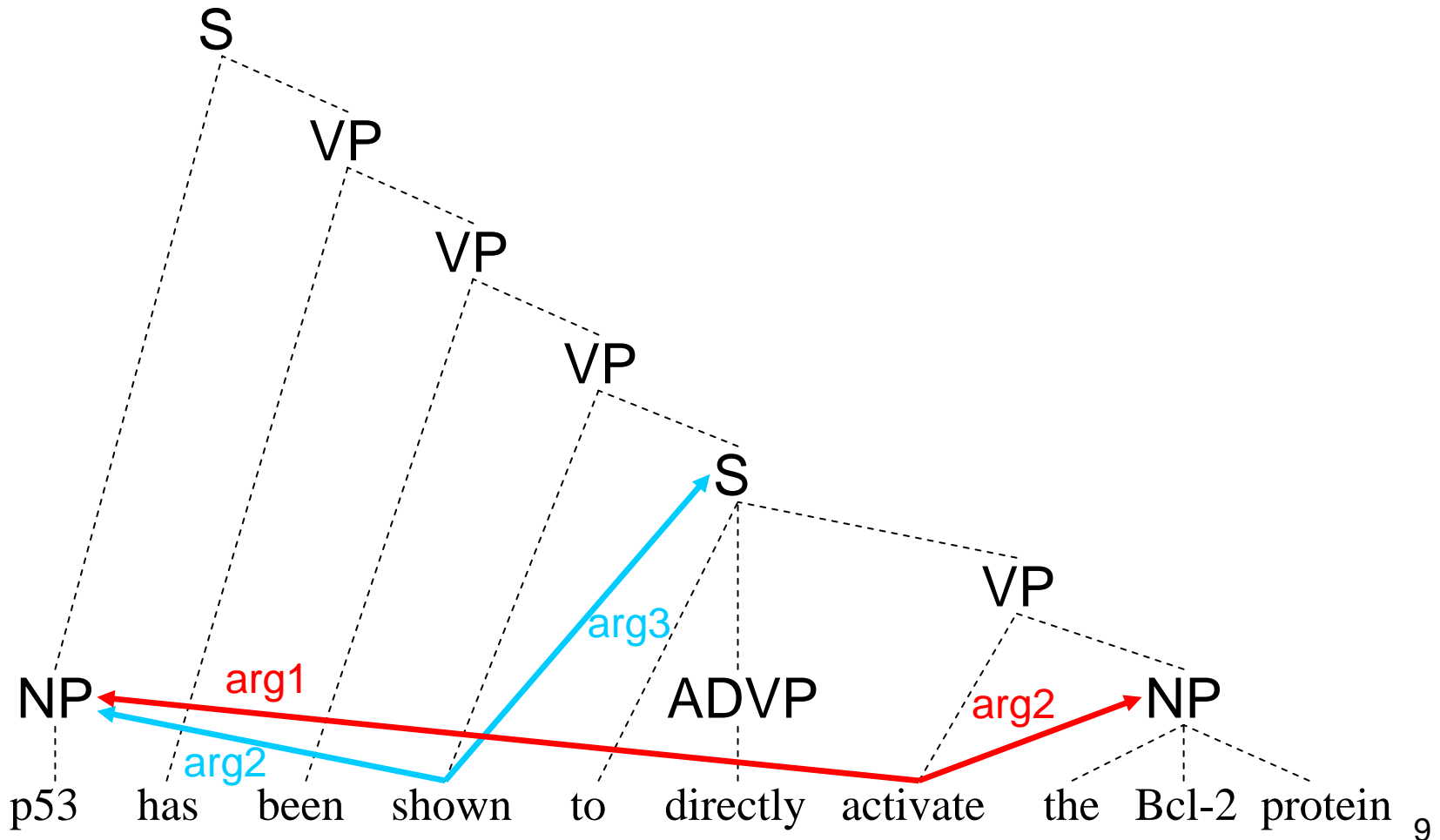
# Demos

- MEDIE
- Info-PubMed
- TerMine



# Predicate-argument structure

Parser based on Probabilistic HPSG (Enju)



# Performance of Semantic Parser

## [Domain Adaptation]



	Penn Treebank	GENIA
Coverage	99.7%	99.2%
F-Value (PA relations)	87.4%	86.4%
Sentence Precision	39.2%	31.8%
Processing Time	0.68sec	1.00sec

# Scalability of TM Tools

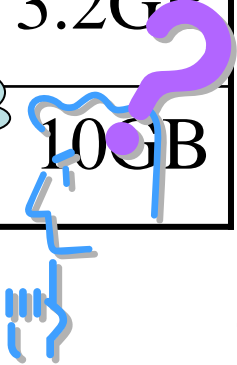
The University of Manchester

Target Corpus: MEDLINE corpus

The number of papers	14,792,890
The number of abstracts	7,434,879
The number of sentences	1,480
The number of words	1,650
Computation time	70 million seconds, that is, about 2 years
Uncompressed size	3.2GB
Compressed size	10GB

Suppose, for example, that it takes one second for parsing one sentence...

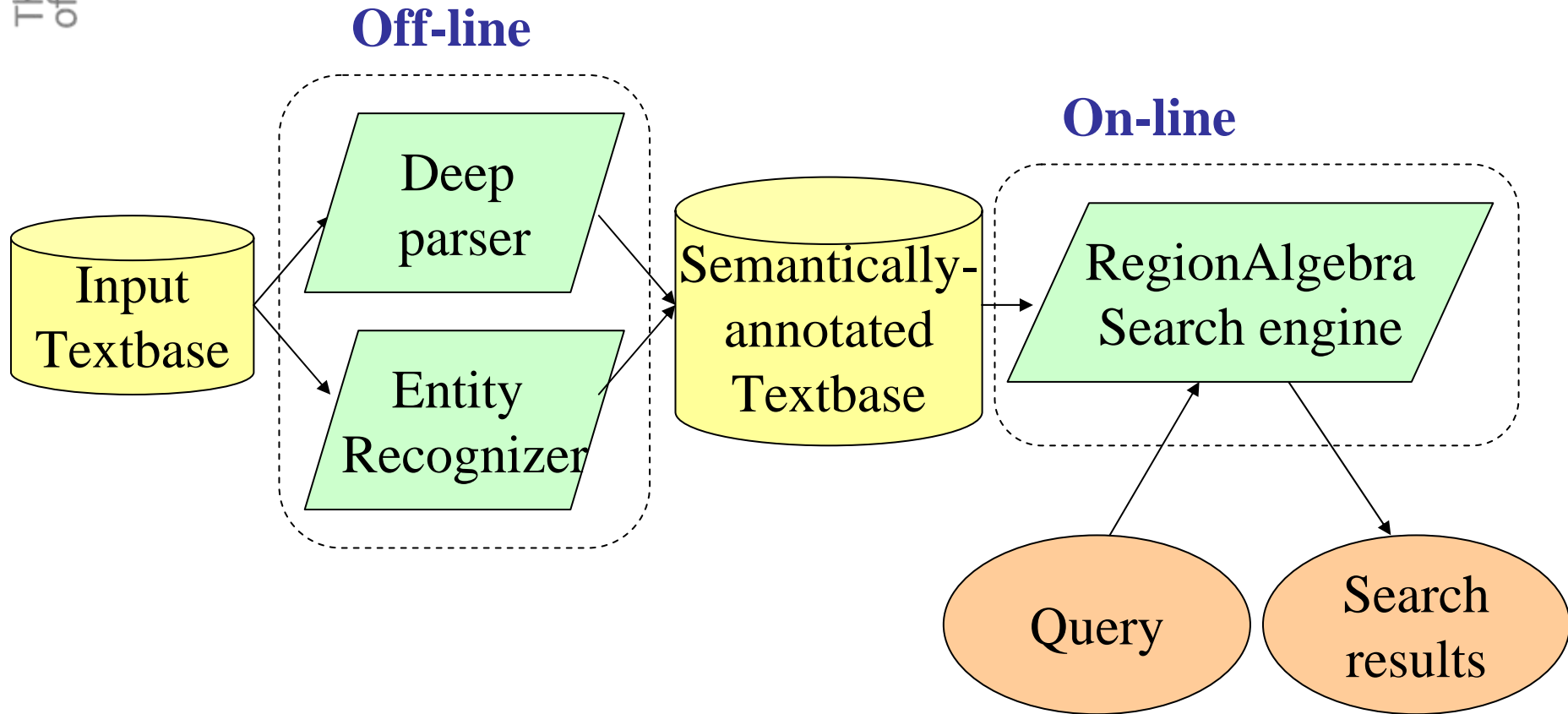
70 million seconds, that is, about 2 years



# TM and GRID

- Solution
  - The entire MEDLINE were parsed by distributed PC clusters consisting of 340 CPUs
  - Parallel processing was managed by grid platform GXP [Taura2004]
- Experiments
  - The entire MEDLINE was parsed in 8 days
- Output
  - Syntactic parse trees and predicate argument structures in XML format
  - The data sizes of compressed/uncompressed output were 42.5GB/260GB.

# Medie system overview



# IMS and TM

## Information Management System

Information in Text is integrated with data and “knowledge” of the domain,  
Persistent text bases  
Intelligent retrieval systems

## Text Mining

On the fly, The results are not shared by a large community



# Demos

- MEDIE
- Info-PubMed
- TerMine

## Managing texts, data representation and their semantics

### Data representation

#### Data Base Module

#### DB of Feature Objects

[content Ubiquitin]

[content [Event  
Pred bind  
agent]]

**Ubiquitin E is bound with**

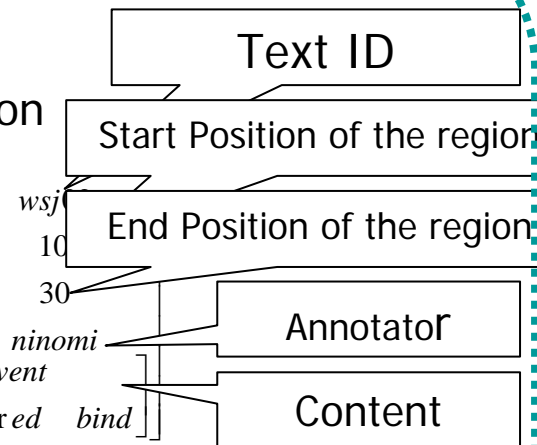
Text DB

Text

### Semantics

#### Copy and Unification

extent  
textid  
startp  
endp  
dc:creator  
content



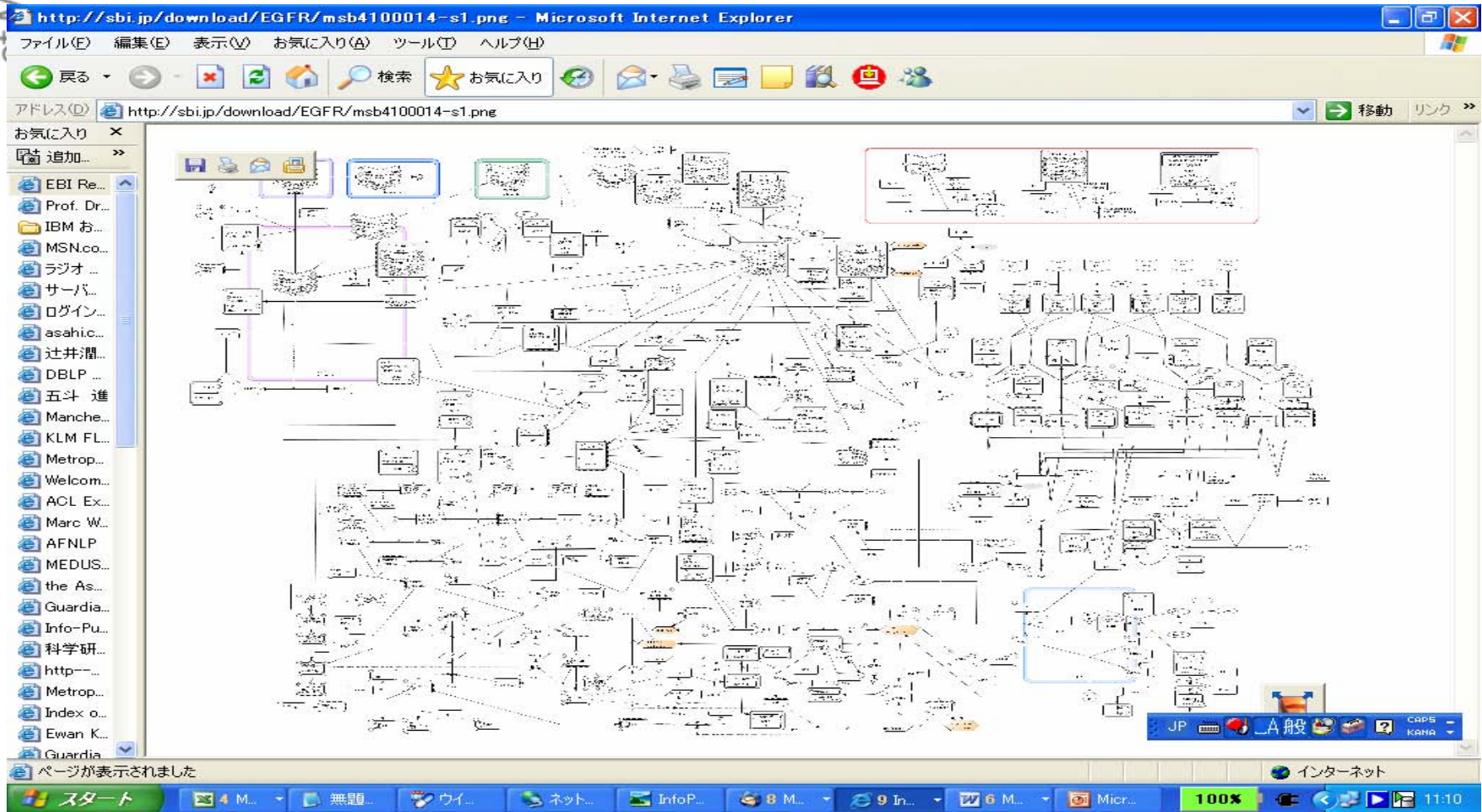
Specialization by unification

extent  
textid wsj02  
startp 10  
endp 30  
dc:creator ninomi  
content [Event  
Pred bind]

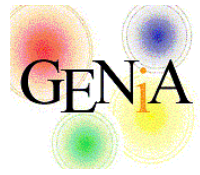
Adding more augmented information induced by inference, type restriction, unification



# Future Plan



Kitano's group, Kell's group



# Future Directions

- Domain Adaptation + Inter-operability
  - High performance can be obtained by using domain specific characteristics and domain semantics
  - Differences among abstracts, full papers, comments in DBs
  - Standardized Interfaces (API) of NLP tools
- Text Archives
  - Abstracts + Full Papers + Comments/Summary Descriptions in DBs
- Combining NLP tools with Mining tools
  - Knowledge Discovery (Disease Gene Association)
  - Hypotheses Generation
  - Automatic Data Interpretation

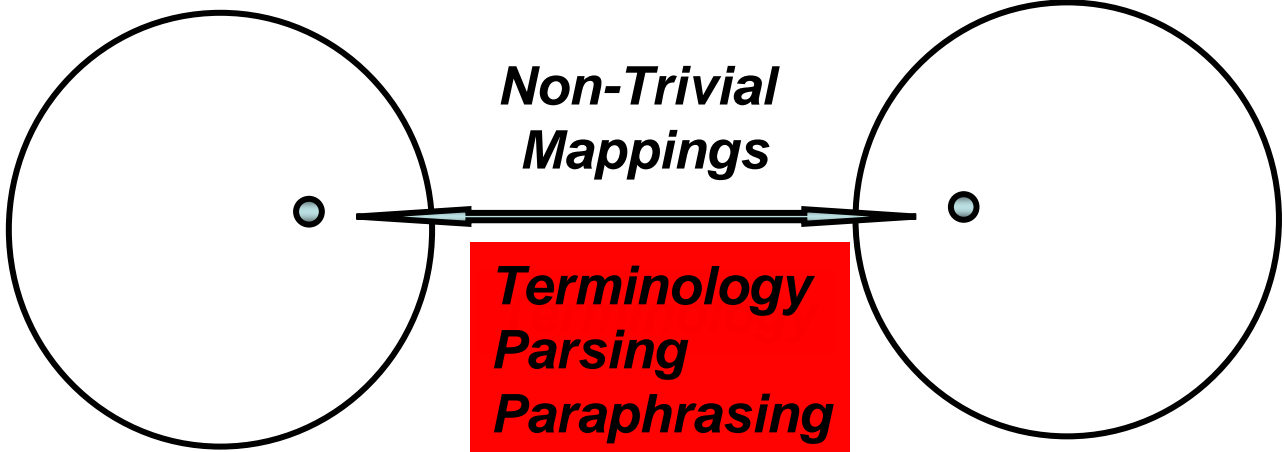
# Plan of the Talk

- Mapping from the LD to KD
  - Terminological Processing
  - Semantic Parsing
- NLP Tools: Domain/Task Adaptation
  - POS Taggers
  - NER
  - Semantic Parsing
- Corpus Building
  - Event Annotation
- Concluding Remarks

# Plan of the Talk

- Mapping from the LD to KD
  - Terminological Processing
  - Semantic Parsing
- NLP Tools: Domain/Task Adaptation
  - POS Taggers
  - NER
  - Semantic Parsing
- Corpus Building
  - Event Annotation
- Concluding Remarks

**From surface diversities and ambiguities  
to  
conceptual invariants**



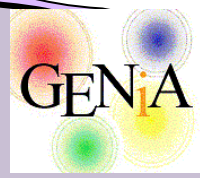
**Language Domain**

**Knowledge Domain**

**Linguistic expressions**

**Concepts and Relationships  
among Them**

**Motivated  
Independently of language**



acronym

Expanded form

Synonym

NF-kappa B  
NF kappa B  
NFKB factor  
NF-KB  
NF kB



nuclear factor-kappa B



Spelling variation

nuclear-factor kappa B  
nuclear factor kappa B  
nuclear factor  $\kappa$  B  
Nuclear Factor kappa B  
.....

# Acronym Dictionary

- N.Okazaki, S.Ananiadou (NaCTeM)
  - Statistics-based Acronym recognizer (not rule-based)

# Enumerating long-form candidates an acronym



- Tokenize a contextual sentence by non-alphanumeric characters (e.g., space, hyphen, etc.)
- Apply Porter's stemming algorithm [Porter 80]
- Extract terms that match the following pattern

[ :WORD: ] . \* \$

We studied the expression of thyroid transcription factor-1 (*TTF-1*).

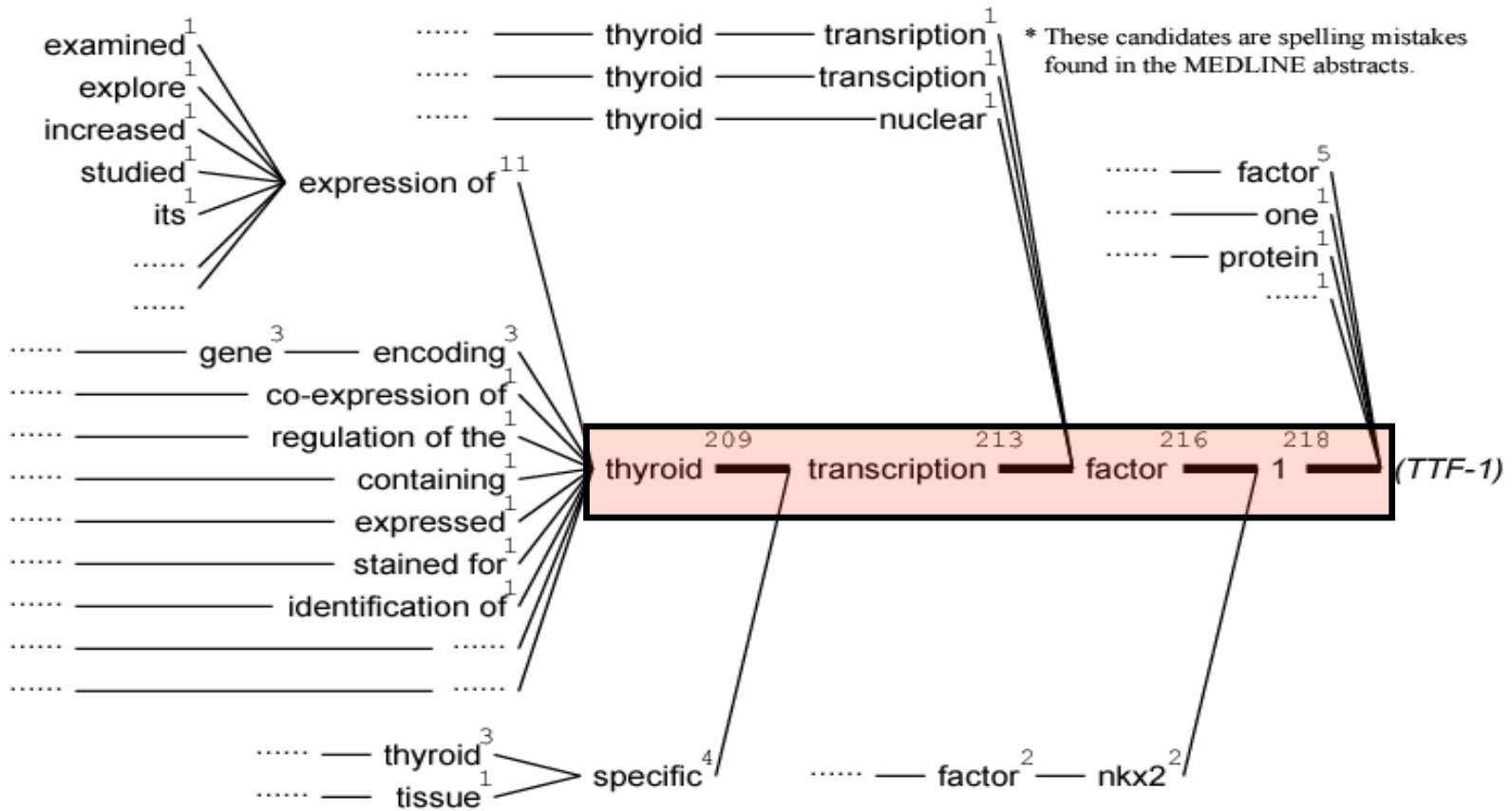
		1
	factor	1
	transcript	factor 1
	thyroid transcript	factor 1
	expression of thyroid transcript	factor 1
studi	the expression of thyroid transcript	factor 1

of thyroid transcript factor 1  
thyroid transcript

Empty string or words of any length



# Expansions for TTF-1



# Results and Demo

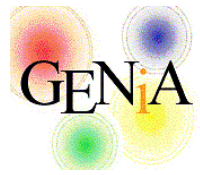
- Term recognition approach to extract acronyms and their definitions from a large text collection
  - usefulness of statistical information for recognizing acronyms
  - 99.1% P and 98.7% R on the evaluation corpus 637,957 contextual sentences, 100 acronyms with 4,024 distinct long forms

Demo of [AcroMine](#)

# Experiment

## [Tsuruoka, et.al. 03 SIGIR]

- Corpus
  - MEDLINE: the largest collection of abstracts in the biomedical domain
- Rule learning
  - 83,142 abstracts
  - Obtained rules: 14,158
- Evaluation
  - 18,930 abstracts
  - Count the occurrences of each generated variant.



1.000	tumor necrosis factor A	0
0.316	TNF A	1
0.200	tumor necrosis factor	1653
0.158	TNF alpha	358
0.133	TNFA	32
0.133	TNF	2631
0.133	Tumour necrosis factor alpha	14
0.133	Tumor Necrosis Factor alpha	2
0.133	Tumor Necrosis Factor-Alpha	0
0.133	TUMOR NECROSIS FACTOR.ALPHA	0
0.133	Tumor necrosis factor alpha	52
0.133	Tumor Necrosis Factor-alpha	8
0.133	TNF-Alpha	0
0.133	TNF-alpha	6899



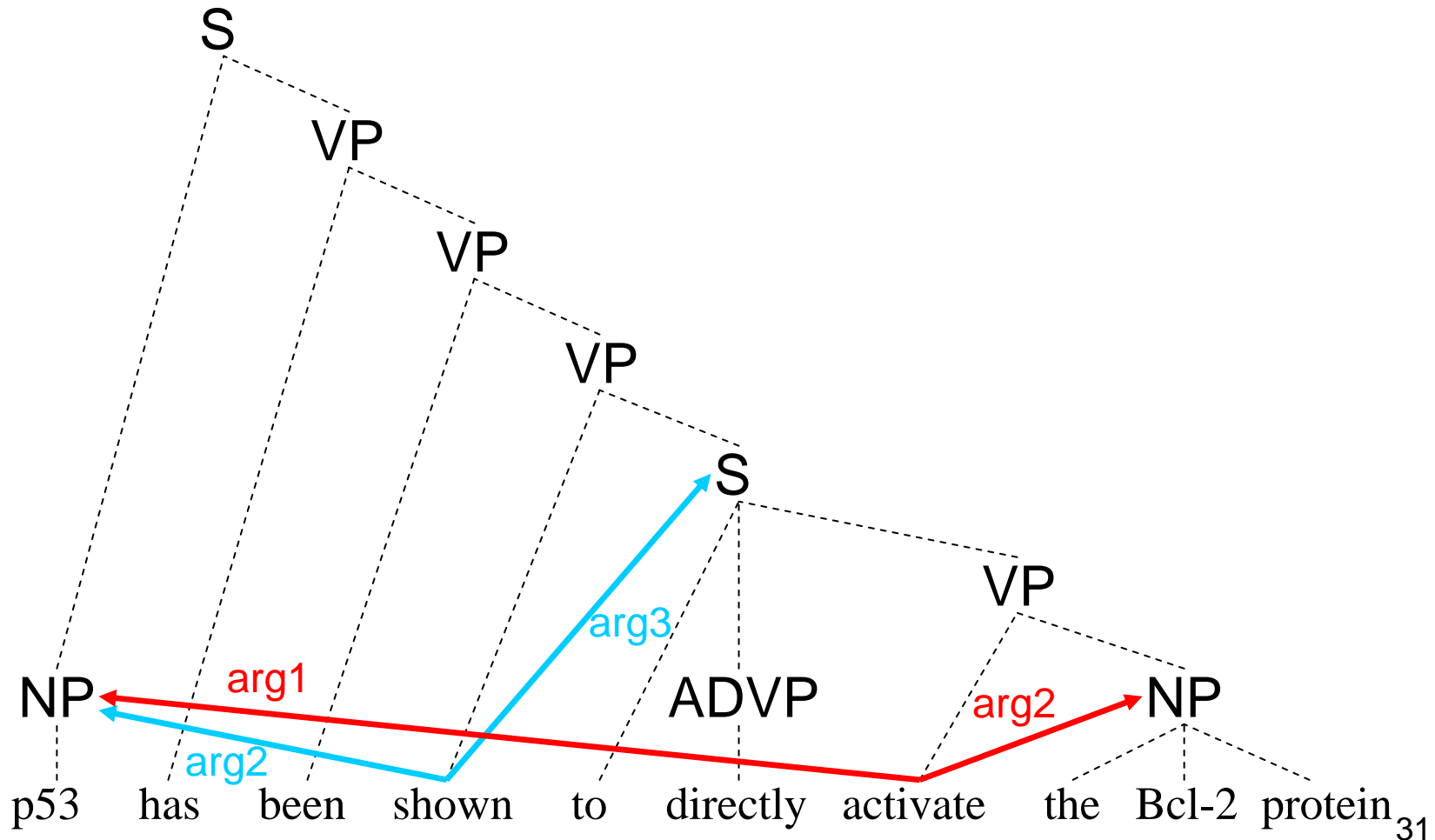
1.000	Human immunodeficiency virus type 1	4
0.400	HIV type 1	52
0.213	HLA - Human immunodeficiency virus type 1	0
0.114	Human immunodeficiency virus type-1	5
0.105	Human immuno virus type 1	0
0.085	HLA - HIV type 1	0
0.077	Human immunodeficiency virus 1	3
0.074	Human immunodeficiency deficiency virus type 1	0
0.059	human immunodeficiency virus type 1	526
0.045	HIV type-1	2
0.041	HLA-- Human immunodeficiency virus type 1	0
0.033	Human immunodeficiency virus type1	0
0.032	HIV-type 1	0
0.032	Human Immunodeficiency virus type 1	0
0.031	HIV 1	8
0.030	HIV-1	6044
0.029	Human immunodeficiency-virus type 1	0
0.028	Human immunodeficiency viru type 1	0
0.025	Human-immunodeficiency virus type 1	0
0.025	HLA - HLA - Human immunodeficiency virus type 1	0

# Plan of the Talk

- Mapping from the LD to KD
  - Terminological Processing
  - **Semantic Parsing**
- NLP Tools: Domain/Task Adaptation
  - POS Taggers
  - NER
  - Semantic Parsing
- Corpus Building
  - Event Annotation
- Concluding Remarks

# Predicate-argument structure

Parser based on Probabilistic HPSG (Enju)



# Performance of Semantic Parser

## [Domain Adaptation]



	Penn Treebank	GENIA
Coverage	99.7%	99.2%
F-Value (PA relations)	87.4%	86.4%
Sentence Precision	39.2%	31.8%
Processing Time	0.68sec	1.00sec



# Scalability of TM Tools

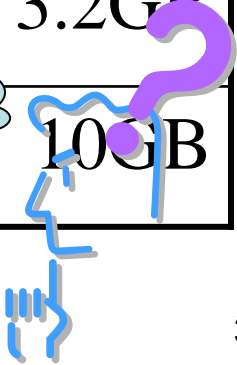
The University of Manchester

Target Corpus: MEDLINE corpus

The number of papers	14,792,890
The number of abstracts	7,434,879
The number of sentences	1,480
The number of words	1,650
Computation time	70 million seconds, that is, about 2 years
Uncompressed size	3.2GB
Compressed size	10GB

Suppose, for example, that it takes one second for parsing one sentence...

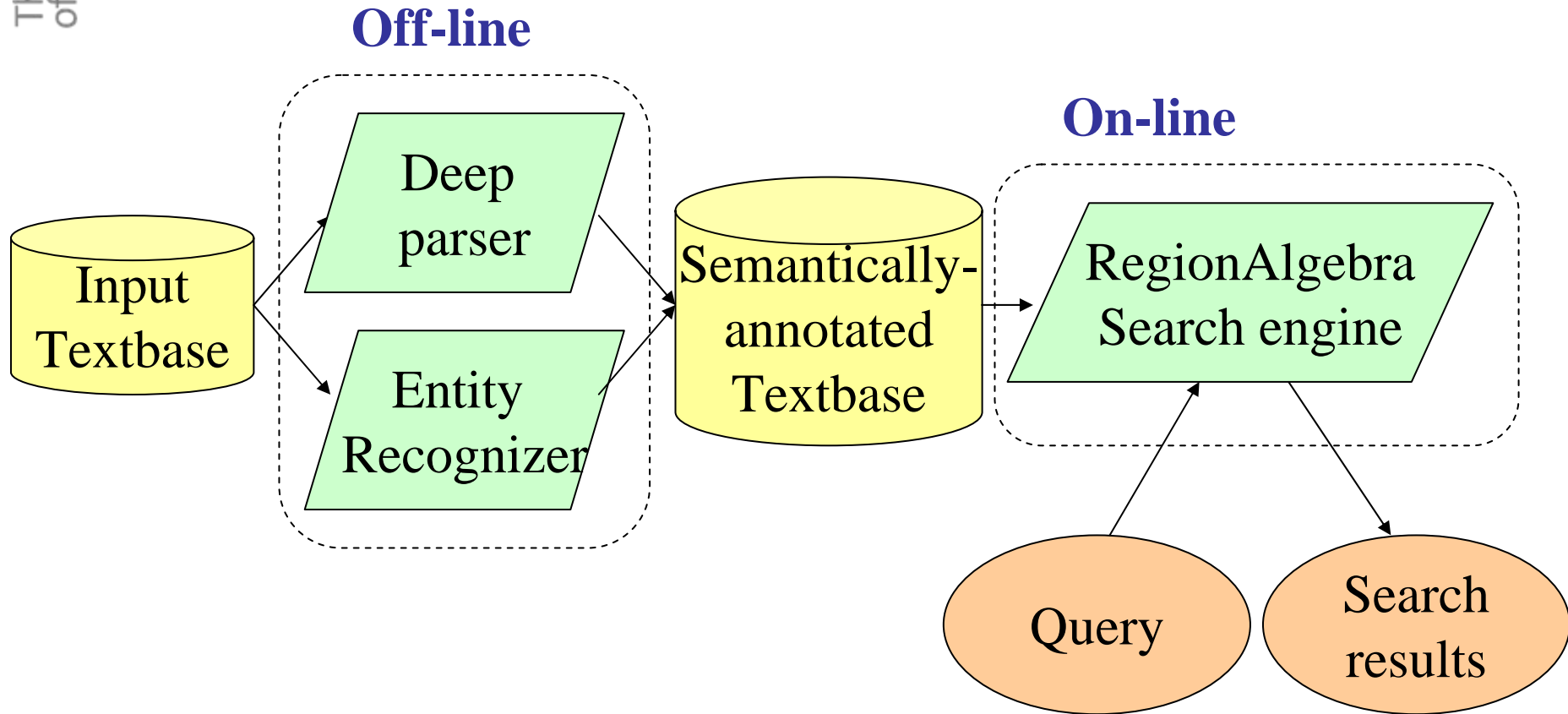
70 million seconds, that is, about 2 years



# TM and GRID

- Solution
  - The entire MEDLINE were parsed by distributed PC clusters consisting of 340 CPUs
  - Parallel processing was managed by grid platform GXP [Taura2004]
- Experiments
  - The entire MEDLINE was parsed in 8 days
- Output
  - Syntactic parse trees and predicate argument structures in XML format
  - The data sizes of compressed/uncompressed output were 42.5GB/260GB.

# Medie system overview



## Managing texts, data representation and their semantics

### Data representation

#### Data Base Module

#### DB of Feature Objects

[content Ubiquitin]

[content [Event  
Pred bind  
agent]]

**Ubiquitin E is bound with**

Text DB

### Text

### Semantics

#### Copy and Unification

```

extent
textid
startp
endp
dc:creator
content
    
```

Text ID

Start Position of the region

End Position of the region

wsj  
10  
30

Annotator

Content



Specialization by unification

```

extent
textid wsj02
startp 10
endp 30
dc:creator ninomi
content [Event  
Pred bind]
    
```

Adding more augmented information induced by inference, type restriction, unification



# Demos

- MEDIE
- Info-PubMed
- TerMine

# Our Policy for IE

- Separate task-independent part from task-specific part.

IE System

PAS = Predicate-Argument Structure

Task-independent

a full parser:

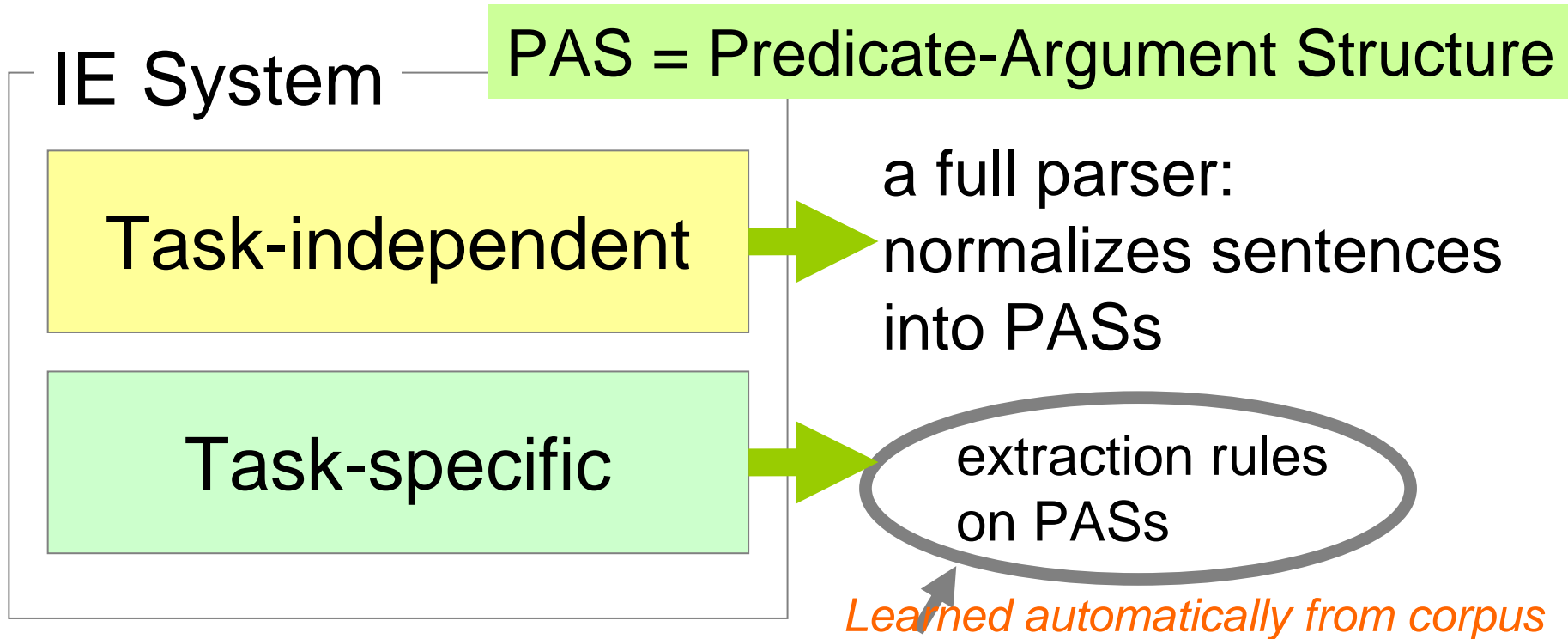
normalizes sentences  
into PASs

Task-specific

extraction rules  
on PASs

# Our Policy for IE

- Distinguish task-independent part from task-specific part.



# Advantages of Full Parsing

- Normalization of syntactic variations into PASs

*Entity1* **activates** *Entity2*

*Entity2* is **activated** by *Entity1*

*Entity1* cooperate to **activate** *Entity2*

*Entity1* play key roles by **activating** *Entity2*

**activate**

ARG1 *Entity1*

ARG2 *Entity2*

We can construct more general extraction rules.

Less extraction rules

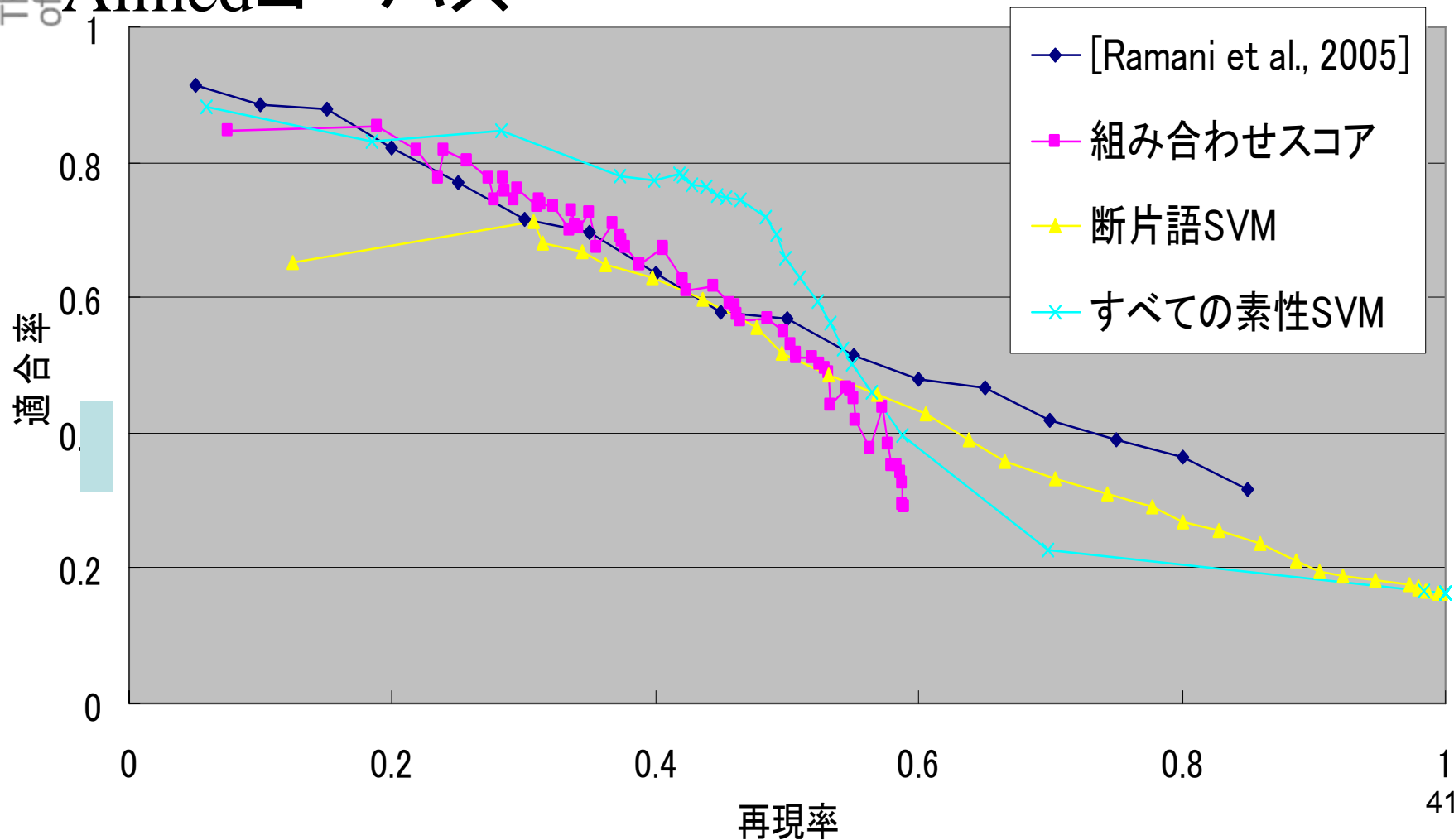


Less training corpora



# Evaluation: Features in SVM

## Aimedコーパス



# Plan of the Talk

- Mapping from the LD to KD
  - Terminological Processing
  - Semantic Parsing
- NLP Tools: Domain Adaptation
  - POS Taggers
  - NER
  - Semantic Parsing
- Corpus Building
  - Event Annotation
- Concluding Remarks

# POS Tagger

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

- General-Purpose POS taggers, trained by WSJ
  - Brill's tagger, TnT tagger, MX POST, etc.
  - 97%
- General-Purpose POS taggers do not work well for MEDLINE abstracts

# Errors seen in TnT tagger (Brants 2000)

A chromosomal translocation in ...

DT JJ NN IN

... 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.

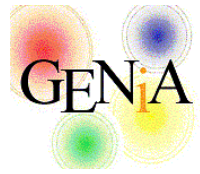
IN VBG DT ~~VBG~~ JJ NN NN

... 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



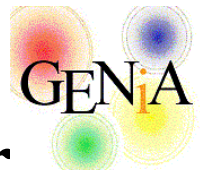
# GENIA tagger

- Probabilistic Model

- Maximum Entropy Markov Model (MEMM)

$$P(t_1^n | w_1^n) = \prod_i P(t_i | t_{i-1} w_1^n)$$

- Maximum Entropy with Inequality Constraints (Kazama and Tsujii 2003)
  - The same effects as Gaussian prior
  - Small number of parameters, Small Models
  - Portable, Efficient, Small size program



# Performance of GENiA Tagger

- GENiA tagger

(Ref.) TnT tagger

Training corpus \	WSJ	GENiA
WSJ	97.0	84.3
GENiA	75.2	98.1
<b>WSJ+GENiA</b>	<b>96.9</b>	<b>98.1</b>

Training corpus \	WSJ	GENiA
WSJ	96.7	84.3
GENiA	80.1	97.9
WSJ+GENiA	96.5	97.5

No degradation of the tagger trained by the mixed corpus

Some degradations (0.2 ~ 0.4) were observed, compared with the taggers trained by “pure” corpora

# Named Entity Recognition (NER)

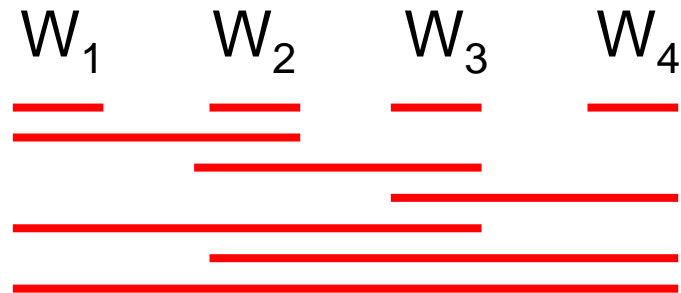
- The first step of IE: to link expressions in text with entities in the knowledge domain, (eg.) person's names with individuals, etc.

“Thus, CIITA not only activates the expression of class II genes but recruits another B cell-specific coactivator to increase transcriptional activity of class II promoters in B cells.”

**PROTEIN** **DNA**

**DNA** **CELL TYPE**

# NER based on segments



- Classification of possible segments in text
  - (Pro) Introduction of features for segments as a whole
  - (Pro) Integration with rule-based systems
  - (Con) Too many segments
- Simple Chunking
  - Exclusion of words that rarely appear in named entities
    - reduction of the number of potential segments



# ML: Maximum Entropy Model

- Features

$$w_{b-2}w_{b-1} = X$$

$$w_{b-2} = X$$

$$w_{b-1}w_{e+1} = X$$

$$w_{b-1} = X$$

$$w_{e+1}w_{e+2} = X$$

$$w_{e+1} = X$$

$$w_b = X$$

$$w_{e+2} = X$$

$$w_e = X$$

$$w_i = X, i \geq b, i \leq e$$

the first and the last letter of  $w_i$  are  
uppercase

$X$  is suffix of  $w_e$ ,  $|X| \leq 5$

- Machine Learner

- Maximum entropy model

- LMVM

- cutoff = 0

- Gaussian prior = 1000

- Training Set:

- 2000 abstracts

- Training set for shared  
task at BioNLP(2004)

# Experiment Results

- Shared task at Coling 2004 BioNLP workshop

	Recall	Precision	F-score
SVM+HMM	76.0	69.4	72.6
<b>Our method</b>	<b>71.5</b>	<b>70.2</b>	<b>70.8</b>
MEMM (Fin 2004)	71.6	68.6	70.1
CRF (Set 2004)	70.3	69.3	69.8

# Performance of Semantic Parser

## [Domain Adaptation]



	Penn Treebank	GENIA
Coverage	99.7%	99.2%
F-Value (PA relations)	87.4%	86.4%
Sentence Precision	39.2%	31.8%
Processing Time	0.68sec	1.00sec

# Reference Distribution

- The statistical model learned by PTB is used as reference distribution
- Estimation of a New Maximum Entropy combined by Reference distribution

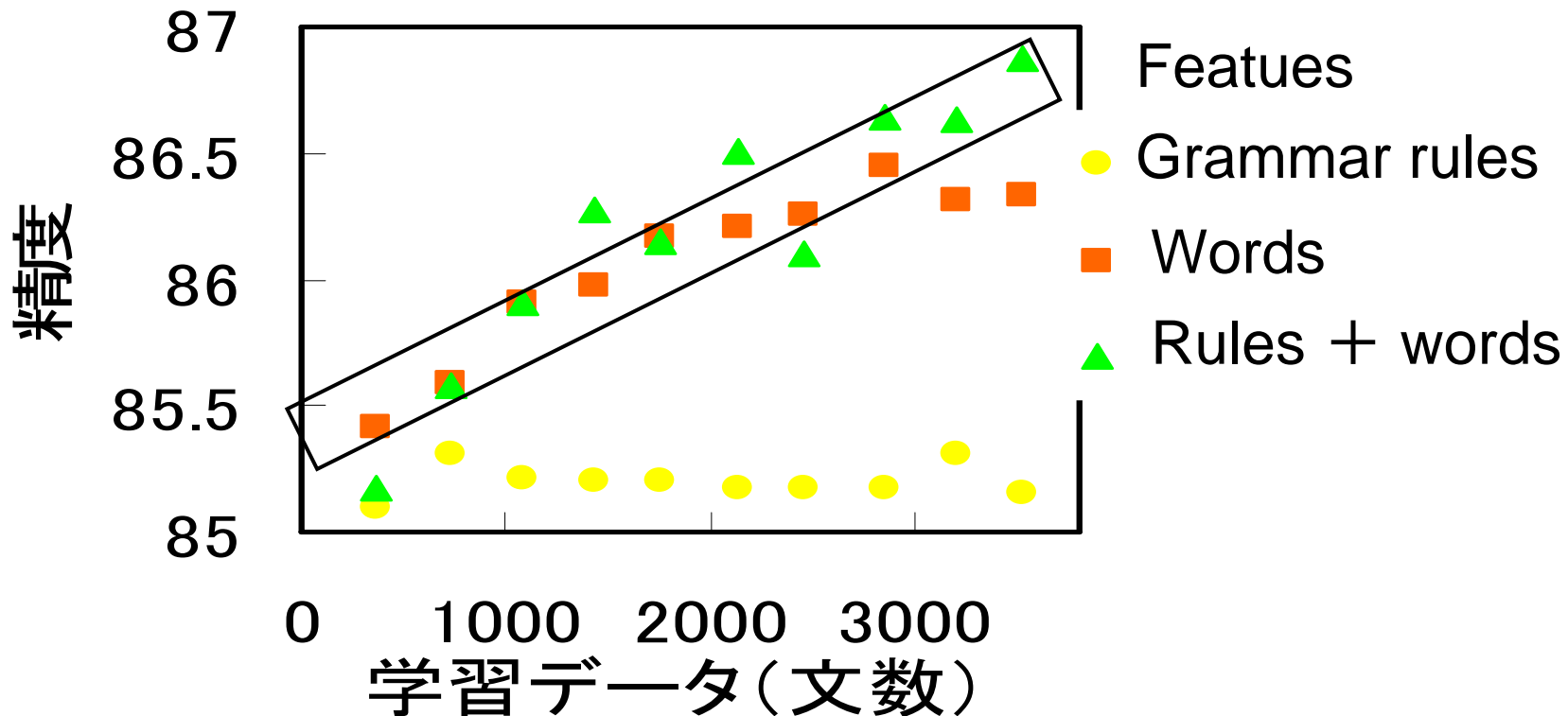
$$p(x|y) = \frac{1}{Z} p_0(x|y) \exp \left( \sum_i \lambda_i f_i(x, y) \right)$$

The diagram illustrates the components of the Maximum Entropy model equation. The term  $p_0(x|y)$  is highlighted in an orange box and labeled "reference distribution". The sum term  $\sum_i \lambda_i f_i(x, y)$  is enclosed in a large bracket, with "Weights" pointing to the  $\lambda_i$  and "features" pointing to the  $f_i(x, y)$ .

- The new model maximizes the model as a whole

# Corpus Size and Precision

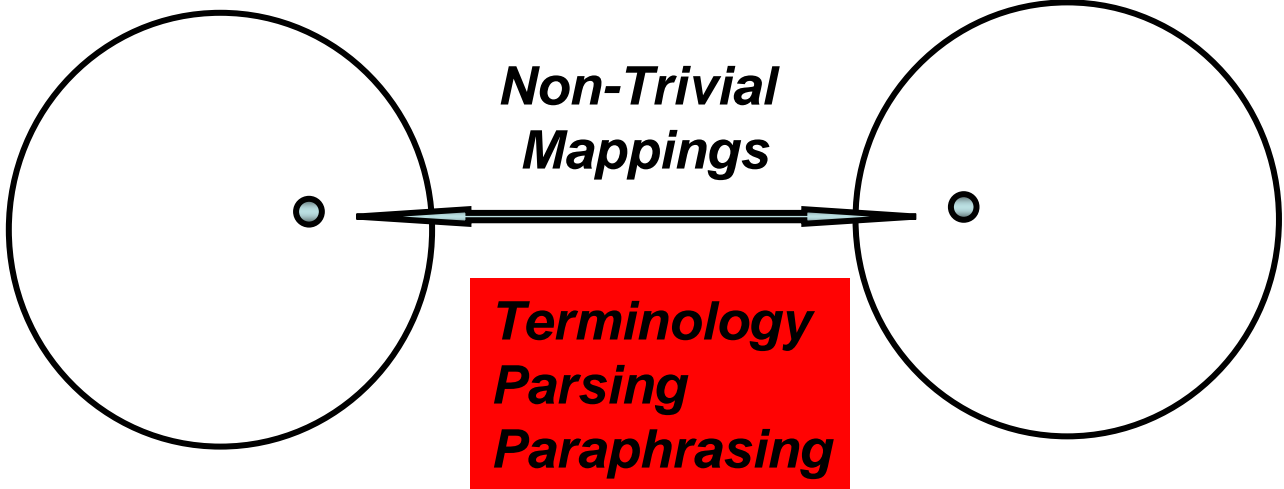
- Domain Adaptation Experiment using GENiA
- Less Learning Time, The Precision improves rapidly



# Plan of the Talk

- Mapping from the LD to KD
  - Terminological Processing
  - Semantic Parsing
- NLP Tools: Domain/Task Adaptation
  - POS Taggers
  - NER
  - Semantic Parsing
- **Corpus Building**
  - **Event Annotation**
- Concluding Remarks

**From surface diversities and ambiguities  
to  
conceptual invariants**



**Language Domain**

**Knowledge Domain**

**Linguistic expressions**

**Concepts and Relationships  
among Them**

**Motivated  
Independently of language**

# Ontology-based Corpus Annotation

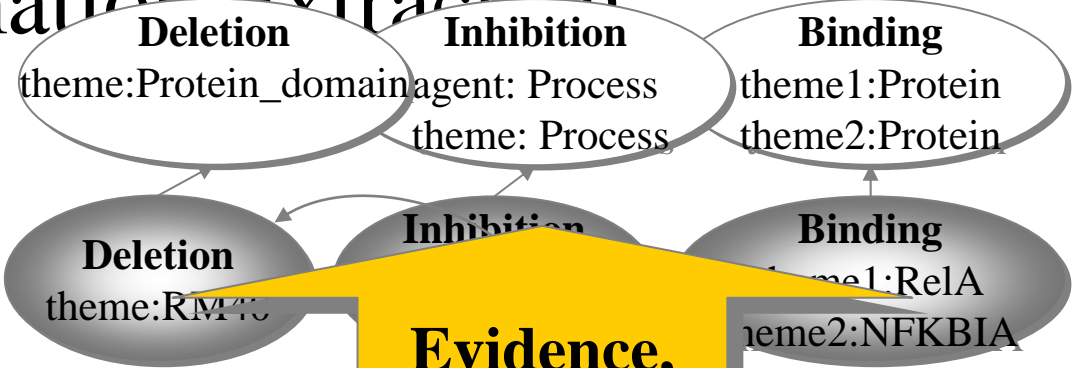
## Information Extraction

ONTOLOGY  
The University  
of Manchester

**contained\_in**  
domain: Protein\_domain  
region: Protein\_domain|Protein

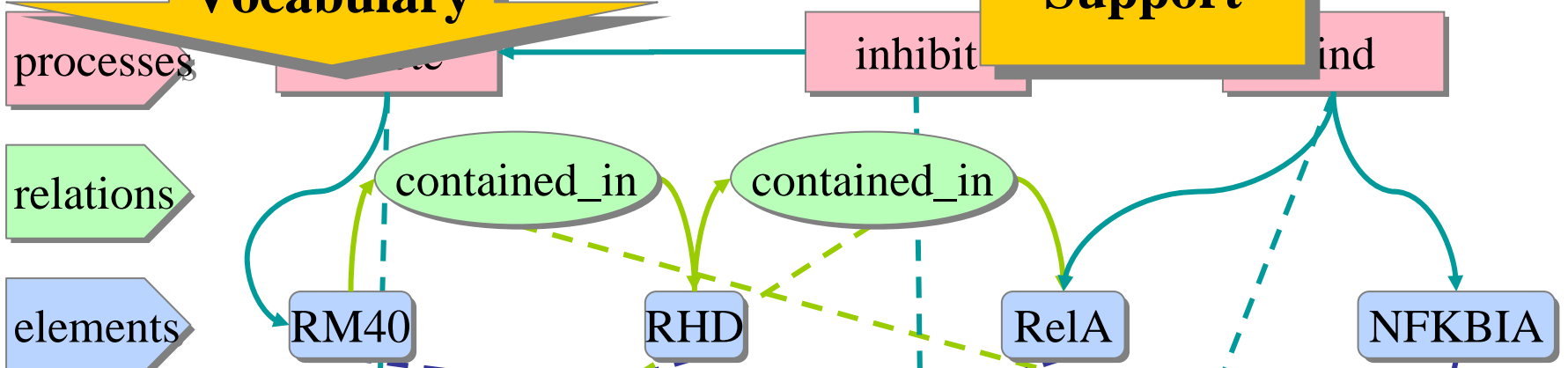


**Domain Knowledge, Vocabulary**



**Evidence, Support**

ANNOTATION  
relations  
elements



TEXT

... 3) selective deletion of the functional nuclear localization signal present in the Rel homology domain of NF-kappa B p65 disrupts its ability to engage I kappa B/MAD-3, and 4) ...

PMID:1493333



# Annotation of GENIA corpus – Term & POS

University  
of  
Manchester

PMID:1984449

Induction of NF-KB during monocyte differentiation by HIV type 1 infection.

PMID:1984449

The pr  
the U9  
purifie  
South  
transa  
long te  
in indu  
macro  
HIV-1  
detect  
was no  
HIV-1

Induction<sub>NN</sub> of<sub>IN</sub> NF-KB<sub>NN</sub> during<sub>IN</sub> monocyte<sub>NN</sub> differentiation<sub>NN</sub> by<sub>IN</sub> HIV<sub>NN</sub> type<sub>NN</sub> 1<sub>CD</sub> infection<sub>NN</sub>.PERIOD

The<sub>DT</sub> production<sub>NN</sub> of<sub>IN</sub> human<sub>U</sub> immunodeficiency<sub>NN</sub> virus<sub>NN</sub> type<sub>NN</sub> 1<sub>CD</sub> (HIV-1<sub>NN</sub>)<sub>PRE</sub> progeny<sub>NN</sub> was<sub>VED</sub> followed<sub>VEN</sub> in<sub>IN</sub> the<sub>DT</sub> U937<sub>NN</sub> promonocytic<sub>U</sub> cell<sub>NN</sub> line<sub>NN</sub> after<sub>IN</sub> stimulation<sub>NN</sub> either<sub>CC</sub> with<sub>IN</sub> retinoic<sub>U</sub> acid<sub>NN</sub> or<sub>CC</sub> PMAN<sub>NN</sub>.COMMA and<sub>CC</sub> in<sub>IN</sub> purified<sub>VEN</sub> human<sub>U</sub> monocytes<sub>NNS</sub> and<sub>CC</sub> macrophages<sub>NNS</sub>.PERIOD Electrophoretic<sub>U</sub> mobility<sub>NN</sub> shift<sub>NN</sub> assays<sub>NNS</sub> and<sub>CC</sub> Southwestern<sub>NN</sub> blotting<sub>NN</sub> experiments<sub>NNS</sub> were<sub>VED</sub> used<sub>VEN</sub> to<sub>TC</sub> detect<sub>VE</sub> the<sub>DT</sub> binding<sub>NN</sub> of<sub>IN</sub> cellular<sub>U</sub> transactivation<sub>NN</sub> factor<sub>NN</sub> NF-KB<sub>NN</sub> to<sub>TC</sub> the<sub>DT</sub> double<sub>U</sub> repeat-KB<sub>U</sub> enhanc<sub>NN</sub> sequence<sub>NN</sub> located<sub>U</sub> in<sub>IN</sub> the<sub>DT</sub> long<sub>U</sub> terminal<sub>U</sub> repeat<sub>NN</sub>.PERIOD PMAN<sub>NN</sub> treatment<sub>NN</sub>.COMMA and<sub>CC</sub> not<sub>RE</sub> retinoic<sub>U</sub> acid<sub>NN</sub> treatment<sub>NN</sub> of<sub>IN</sub> the<sub>DT</sub> U937<sub>NN</sub> cells<sub>NNS</sub> acts<sub>VEZ</sub> in<sub>IN</sub> inducing<sub>VEG</sub> NF-KB<sub>NN</sub> expression<sub>NN</sub> in<sub>IN</sub> the<sub>DT</sub> nuclei<sub>NNS</sub>.PERIOD In<sub>IN</sub> nuclear<sub>U</sub> extracts<sub>NNS</sub> from<sub>IN</sub> monocytes<sub>NNS</sub> or<sub>CC</sub> macrophages<sub>NNS</sub>.COMMA induction<sub>NN</sub> of<sub>IN</sub> NF-KB<sub>NN</sub> occurred<sub>VED</sub> only<sub>RE</sub> if<sub>IN</sub> the<sub>DT</sub> cells<sub>NNS</sub> were<sub>VED</sub> previously<sub>RE</sub> infected<sub>VEN</sub> with<sub>IN</sub> HIV-1<sub>NN</sub>.PERIOD When<sub>WERE</sub> U937<sub>NN</sub> cells<sub>NNS</sub> were<sub>VED</sub> infected<sub>VEN</sub> with<sub>IN</sub> HIV-1<sub>NN</sub>.COMMA not

repeat-KB<sub>U</sub> enhanc<sub>NN</sub> sequence<sub>NN</sub> located<sub>U</sub> in<sub>IN</sub> the<sub>DT</sub> long<sub>U</sub> terminal<sub>U</sub> repeat<sub>NN</sub>.PERIOD PMAN<sub>NN</sub> treatment<sub>NN</sub>.COMMA and<sub>CC</sub> not<sub>RE</sub> retinoic<sub>U</sub> acid<sub>NN</sub> treatment<sub>NN</sub> of<sub>IN</sub> the<sub>DT</sub> U937<sub>NN</sub> cells<sub>NNS</sub> acts<sub>VEZ</sub> in<sub>IN</sub> inducing<sub>VEG</sub> NF-KB<sub>NN</sub> expression<sub>NN</sub> in<sub>IN</sub> the<sub>DT</sub> nuclei<sub>NNS</sub>.PERIOD In<sub>IN</sub> nuclear<sub>U</sub> extracts<sub>NNS</sub> from<sub>IN</sub> monocytes<sub>NNS</sub> or<sub>CC</sub> macrophages<sub>NNS</sub>.COMMA induction<sub>NN</sub> of<sub>IN</sub> NF-KB<sub>NN</sub> occurred<sub>VED</sub> only<sub>RE</sub> if<sub>IN</sub> the<sub>DT</sub> cells<sub>NNS</sub> were<sub>VED</sub> previously<sub>RE</sub> infected<sub>VEN</sub> with<sub>IN</sub> HIV-1<sub>NN</sub>.PERIOD

Ter  
an  
20  
abstracts

# Annotation of GENIA corpus – Process&Tree

**Tree  
annotation  
2000  
abstracts**

PMID:1984449

NP

NP Induction

PP of

NP NF-KB

PP during

NP monocyte differentiation

PP by

NP HIV type 1 infection

PMID:1984449

Induction of NF-KB during monocyte differentiation by HIV type 1 infection.

**INFECTION P1**

THEME: [T9]

PREDICATE: Induction of NF-KB during monocyte differentiation by HIV type 1 infection.

**INDUCTION P2**

AGENT: [I1]

THEME: [P1]

PREDICATE: Induction of NF-KB during monocyte differentiation by HIV type 1 infection.

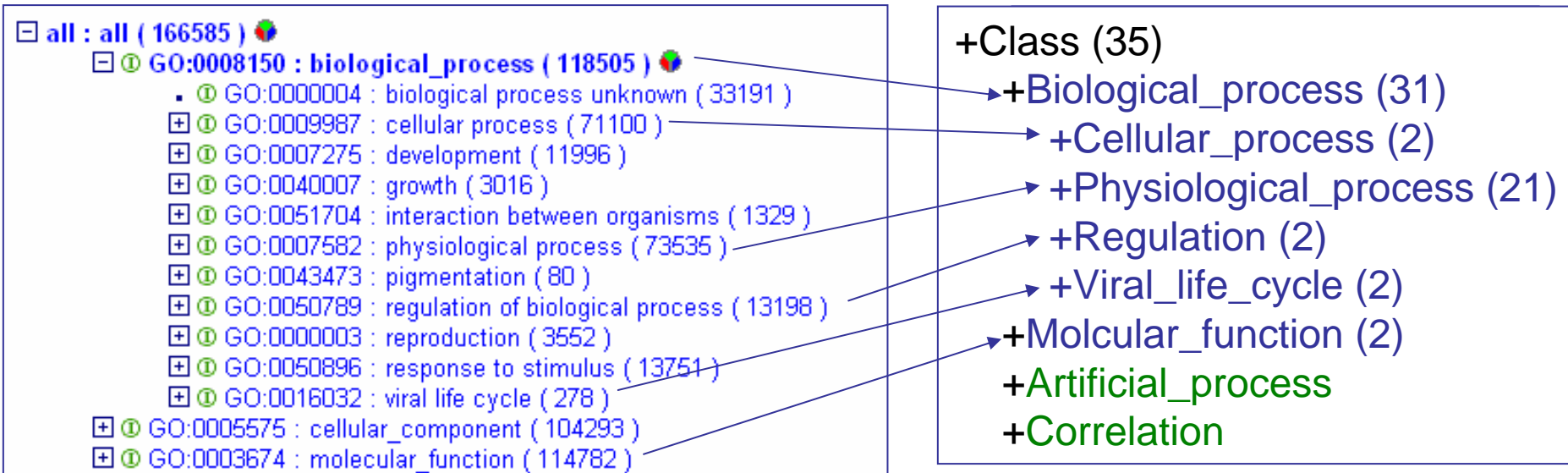
The production of human immunodeficiency virus type 1 (HIV-1) progeny was

**Process  
annotation  
500 abstracts  
(preliminary**

# GENIA event ontology

## Gene ontology

## GENIA ontology



# Plan of the Talk

- Mapping from the LD to KD
  - Terminological Processing
  - Semantic Parsing
- NLP Tools: Domain/Task Adaptation
  - POS Taggers
  - NER
  - Semantic Parsing
- Corpus Building
  - Event Annotation
- Concluding Remarks

# NLP and TM

## Natural language processing

Language as a complex system linking surface strings of characters with their meanings  
Text and words as structured objects



NLP-based TM

## Text Mining

Text as a bag of words  
Words as surface strings



# Future Directions

- Domain Adaptation + Inter-operability
  - High performance can be obtained by using domain specific characteristics and domain semantics
  - Differences among abstracts, full papers, comments in DBs
  - Standardized Interfaces (API) of NLP tools
- Text Archives
  - Abstracts + Full Papers + Comments in DBs
- Combining NLP tools with Mining tools
  - Knowledge Discovery (Disease Gene Association)
  - Hypotheses Generation
  - Automatic Data Interpretation