MANCHESTER 1824

The University of Manchester

GENIA

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

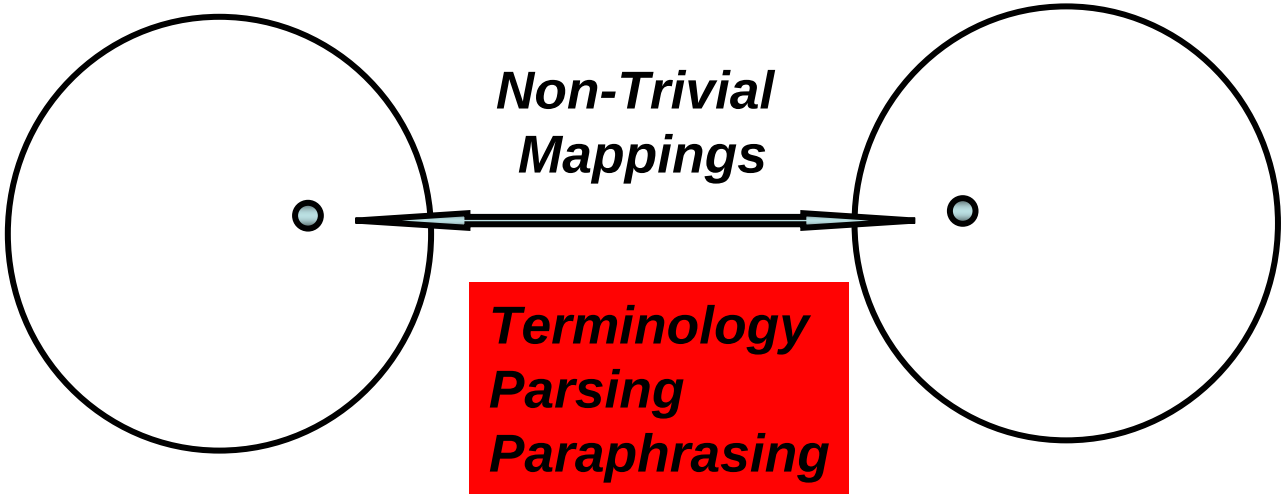
Text Mining

Text as a bag of words

Words as surface strings

NLP-based TM

From surface diversities and ambiguities to conceptual invariants



Language Domain

Linguistic expressions

Knowledge Domain

Concepts and Relationships among Them

Motivated Independently of language

MANCHESTER 1824

The University of Manchester

GENIA

Example

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

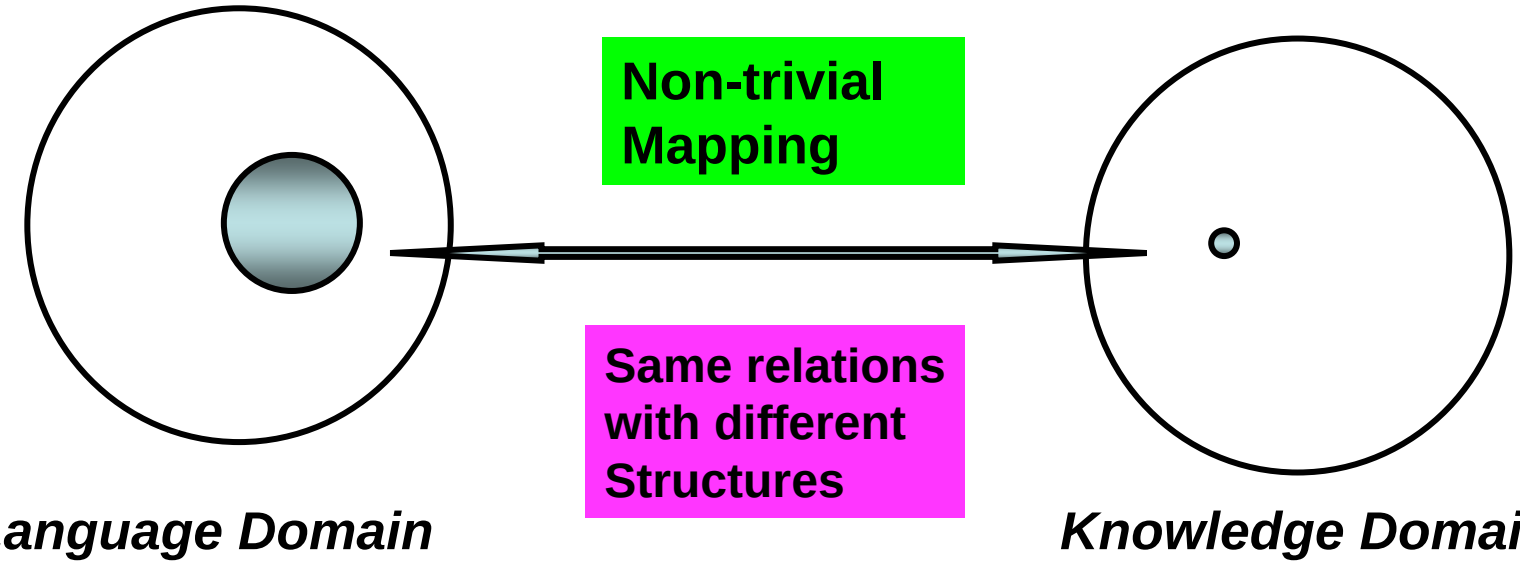
Transcription initiation by the sigma(54)-RNA polymerase holoenzyme requires **an enhancer-binding protein** that is thought to contact sigma(54) to **activate** transcription.

Full-stringe Strauter protein ...r..., we postulate that only phosphorylated **PHO2 protein could activate** the transcription of PHO5 gene.

with oskr mRNA and activate its translation, but fails to ass...

[sentence] > ([arg1_activate] > [protein])

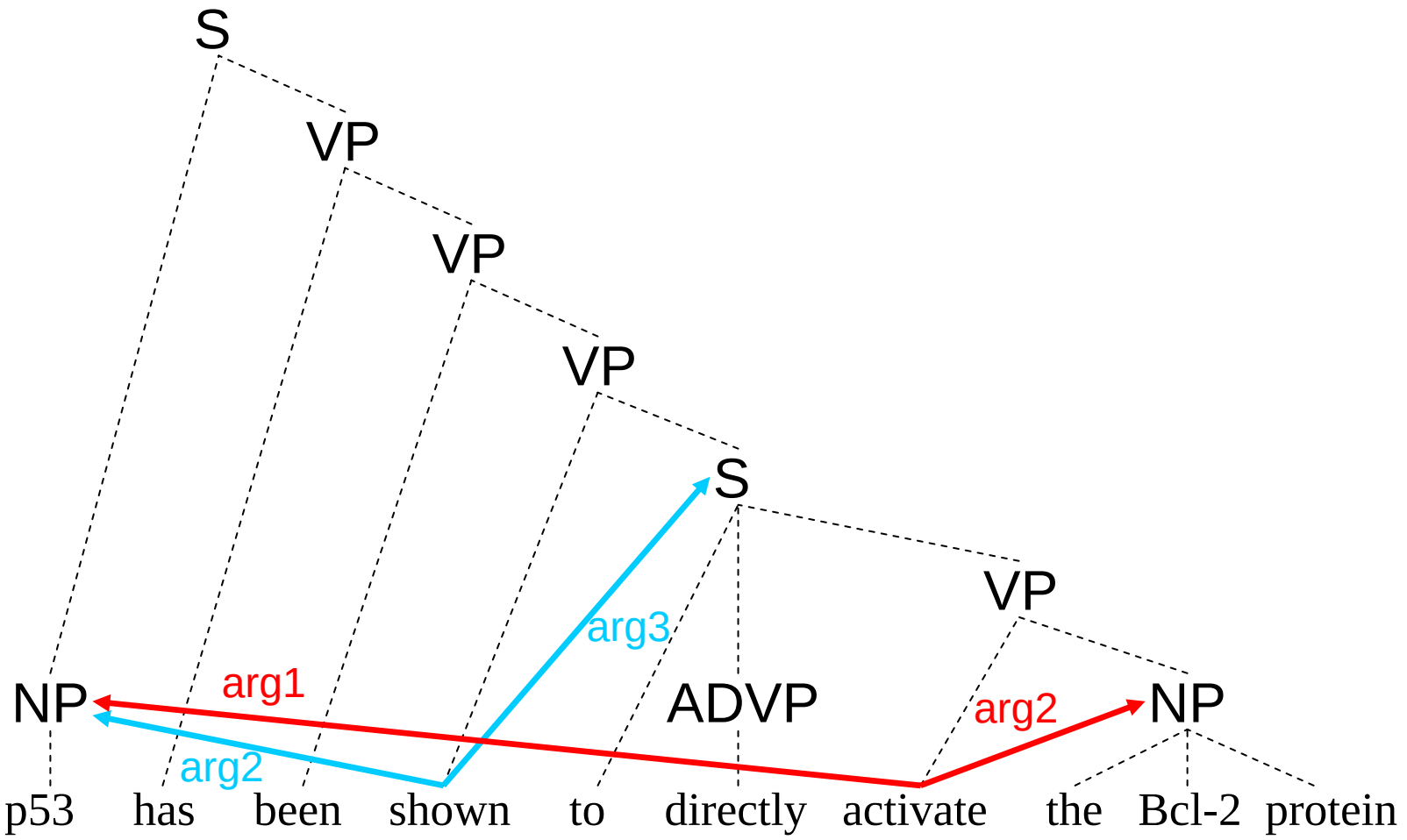
Retrieval using Regional Algebra

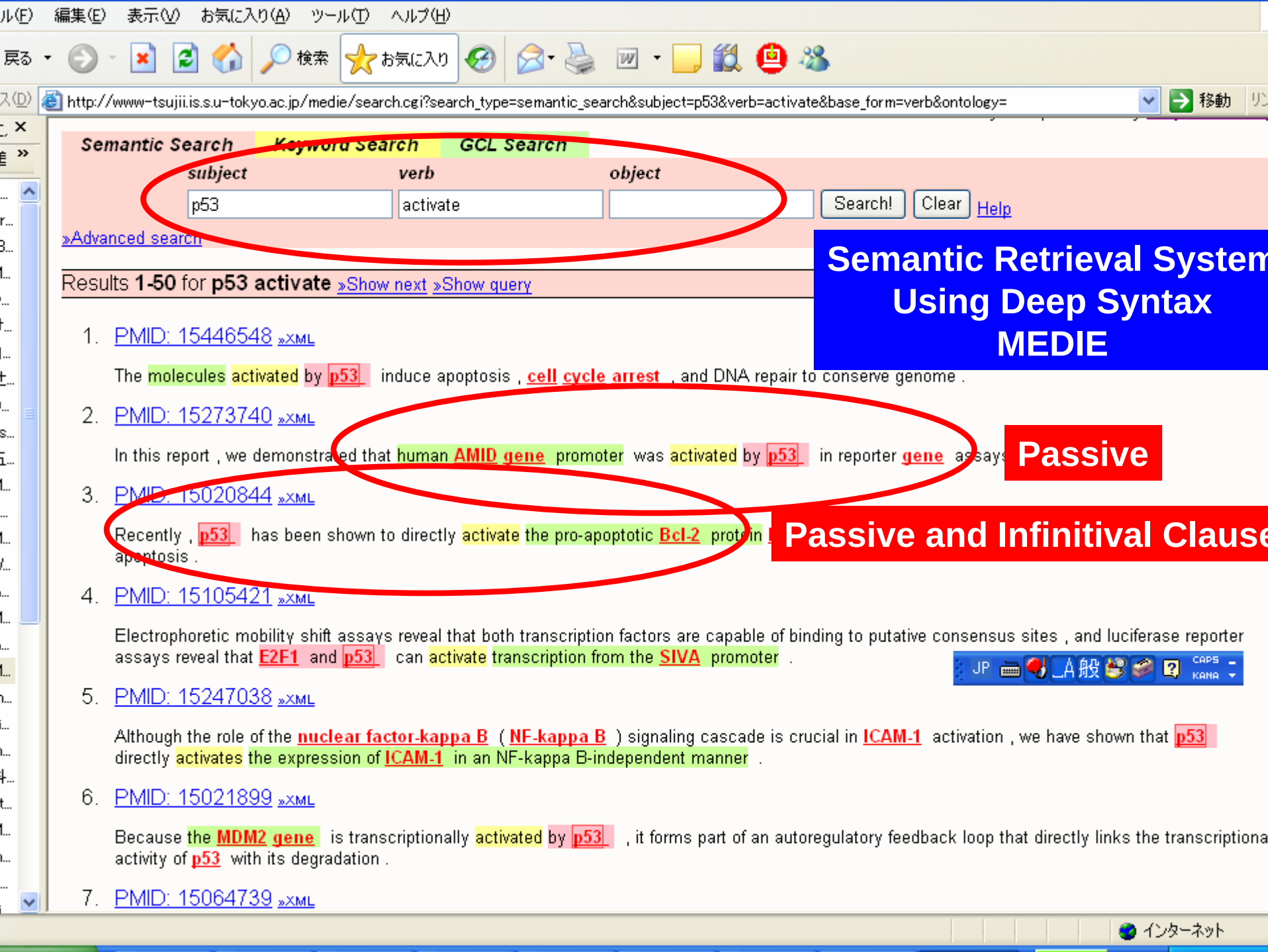


Independently motivated of Language

Predicate-argument structure

Parser based on Probabilistic HPSG (Enju)



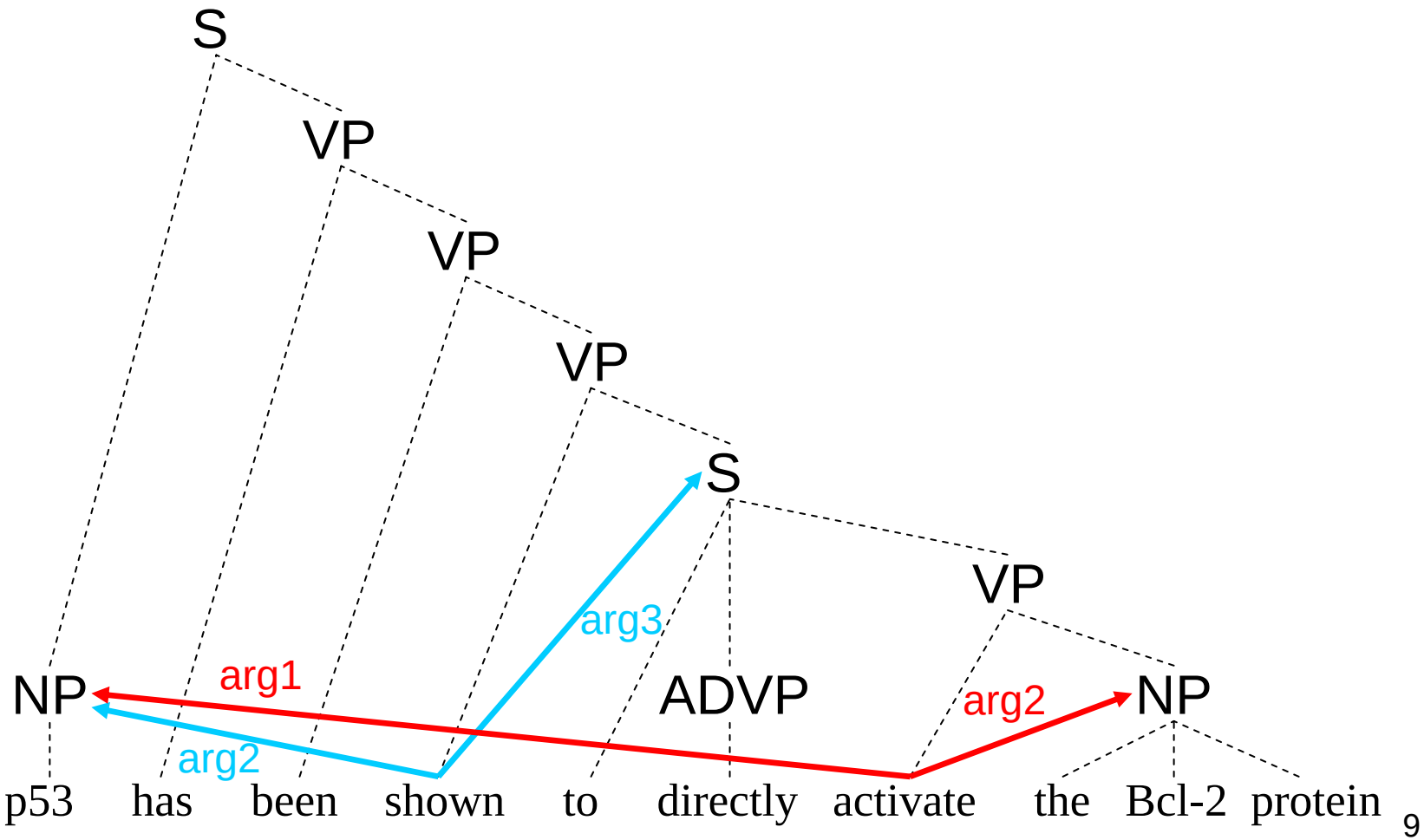

Semantic Search
Keyword Search
GCL Search
subject
verb
object
p53
activate
Search!
Clear
Help
»Advanced search
Semantic Retrieval System
Using Deep Syntax
MEDIE
Results 1-50 for p53 activate »Show next »Show query
1. PMID: 15446548 »XML
The molecules activated by p53 induce apoptosis , cell cycle arrest , and DNA repair to conserve genome .
2. PMID: 15273740 »XML
In this report , we demonstrated that human AMID gene promoter was activated by p53 in reporter gene assays
Passive
3. PMID: 15020844 »XML
Recently , p53 has been shown to directly activate the pro-apoptotic Bcl-2 protein
apoptosis .
Passive and Infinitival Clause
4. 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 .
5. 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 .
6. PMID: 15021899 »XML
Because the MDM2 gene is transcriptionally activated by p53 , it forms part of an autoregulatory feedback loop that directly links the transcriptiona
activity of p53 with its degradation .
7. PMID: 15064739 »XML

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 Precison	39.2%	31.8%
Processing Time	0.68sec	1.00sec

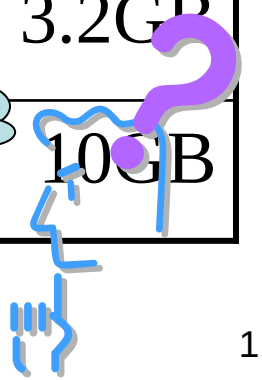
Scalability of TM Tools

Target Corpus: MEDLINE corpus

The number of papers	14,792,890
The number of abstracts	7,434,879
The number of ser	,480
The number of y	,650
Comp	3.2GB
Un	10GB

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

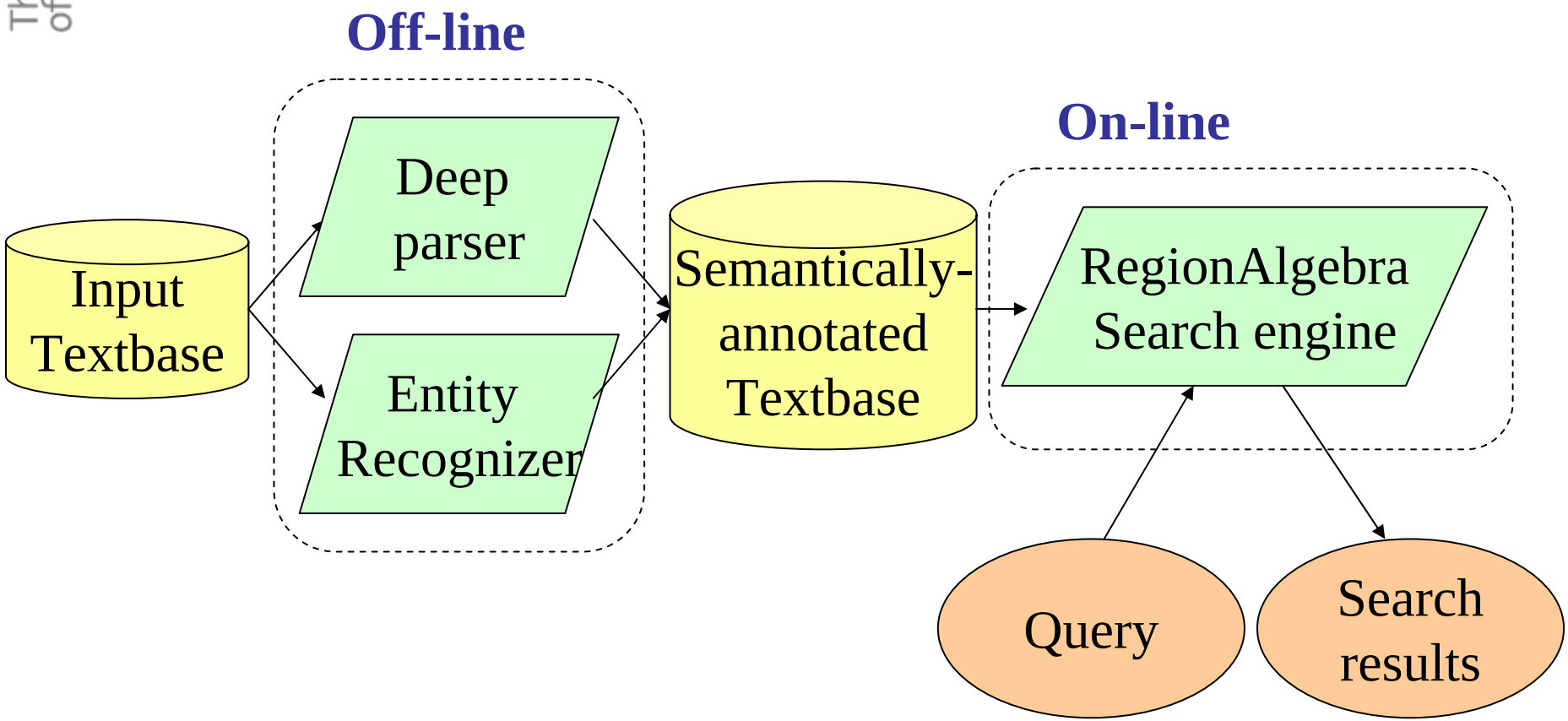
70 million seconds, that is, about 2 years



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

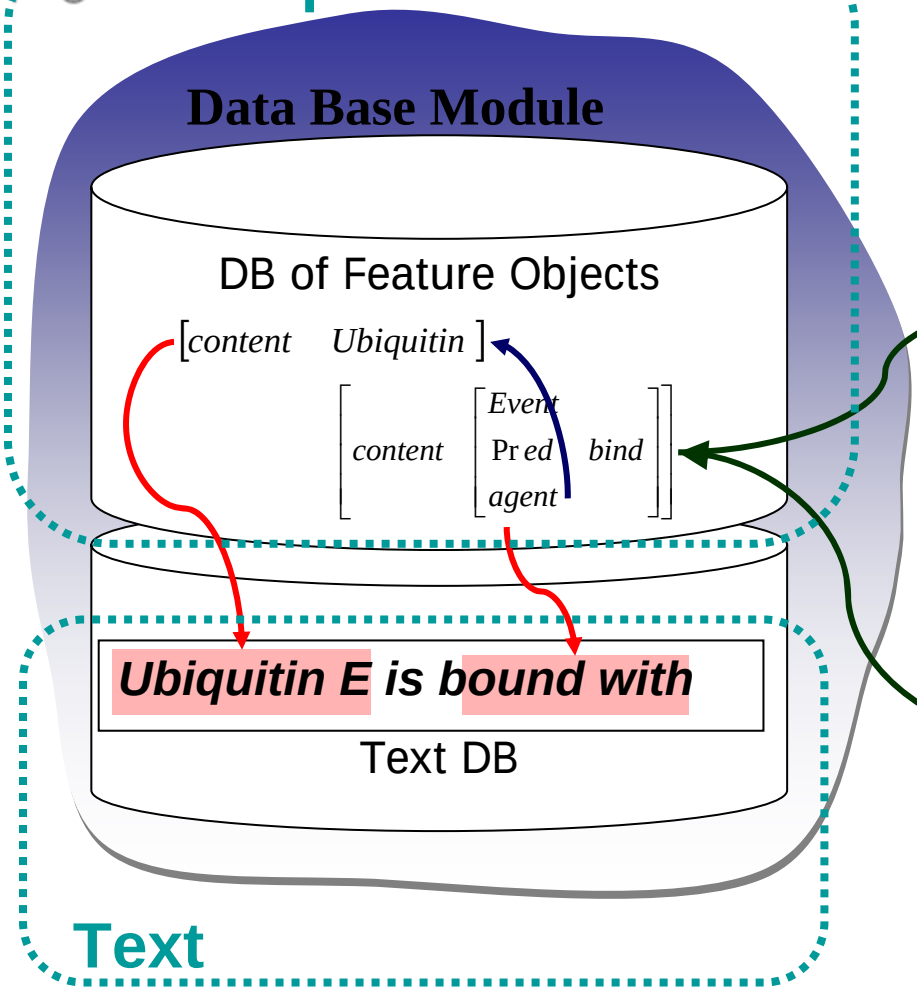
Demos

- MEDIE
- Info-PubMed
- TerMine

Text Archive with Feature Obejcts

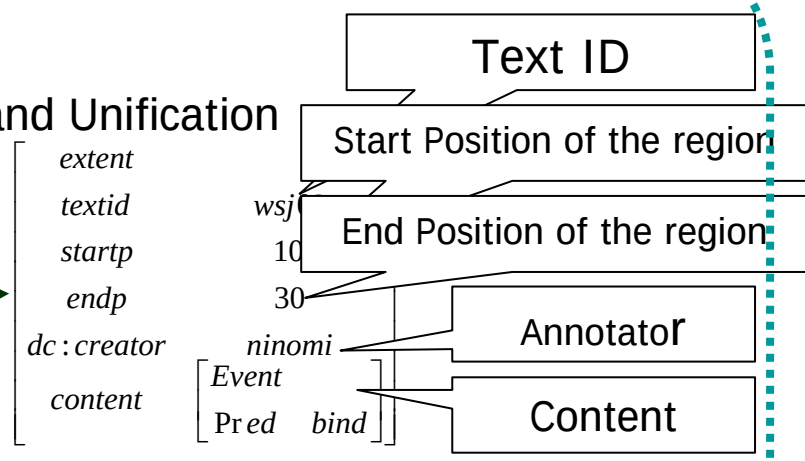
Managing texts, data representation and their semantics

Data representation

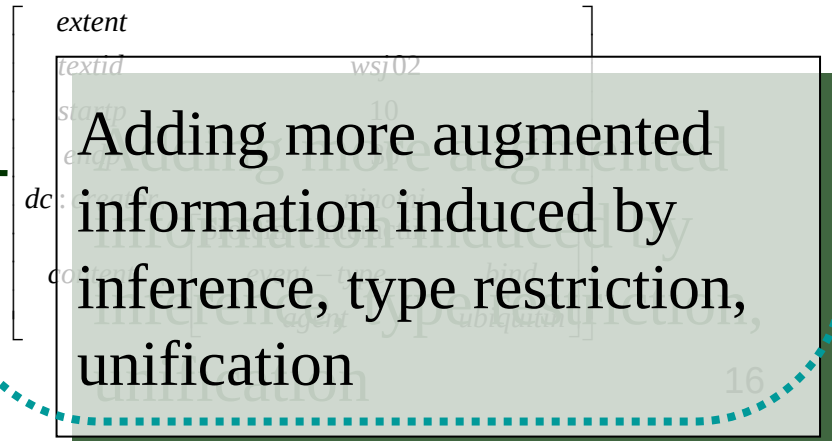


Semantics

Copy and Unification



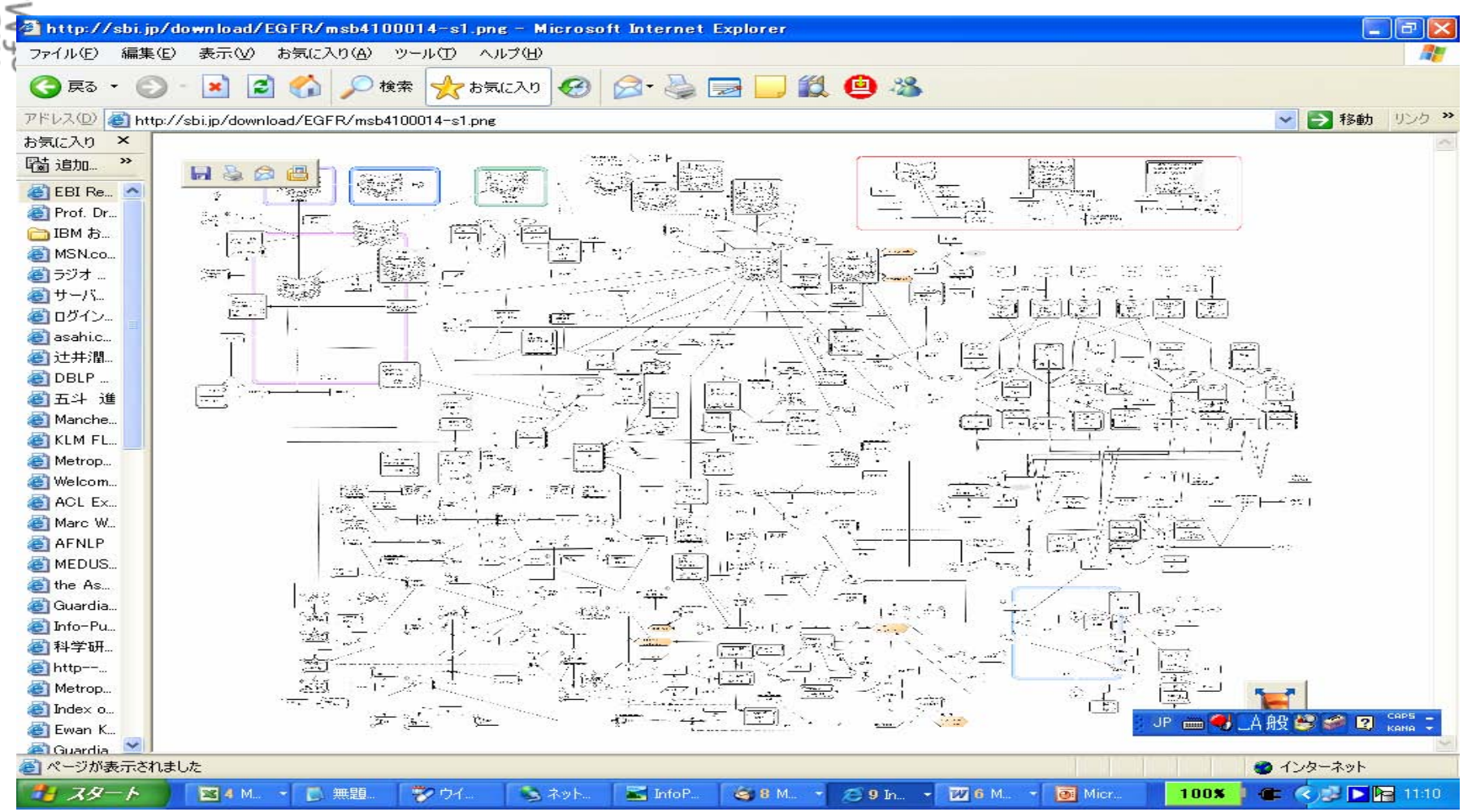
Specialization by unification



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

Text

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

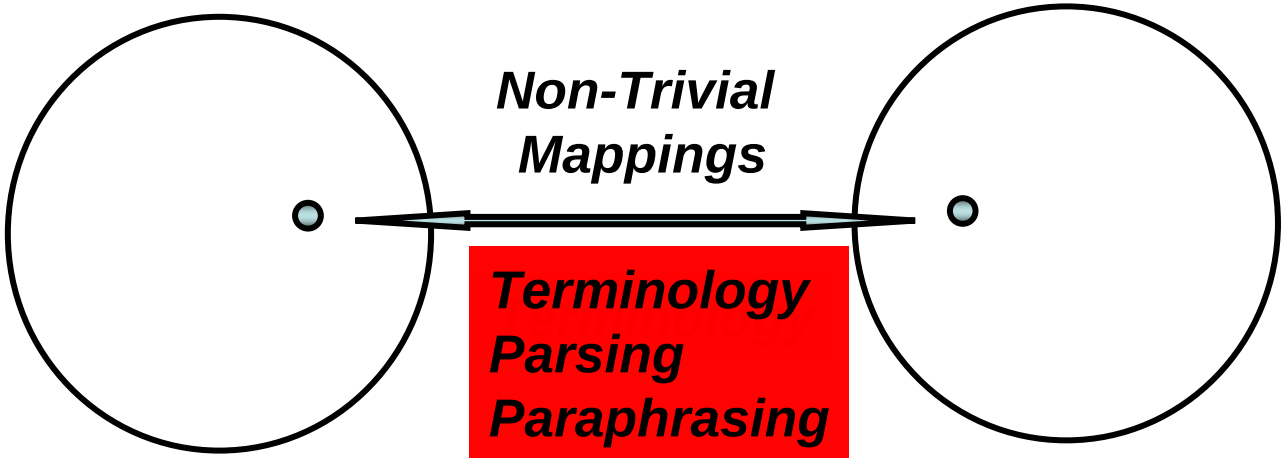
- Mapping from the LD to KD
  - Terminological Processing
  - Semantic Parsing
- NLP Tools: Domain/Task Adaptation
  - POS Taggers
  - NER
  - Semantic Parsing
- Corpus Building
  - Event Annotation
- Concluding Remarks

MANCHESTER 1824

The University of Manchester

GENIA

From surface diversities and ambiguities to conceptual invariants



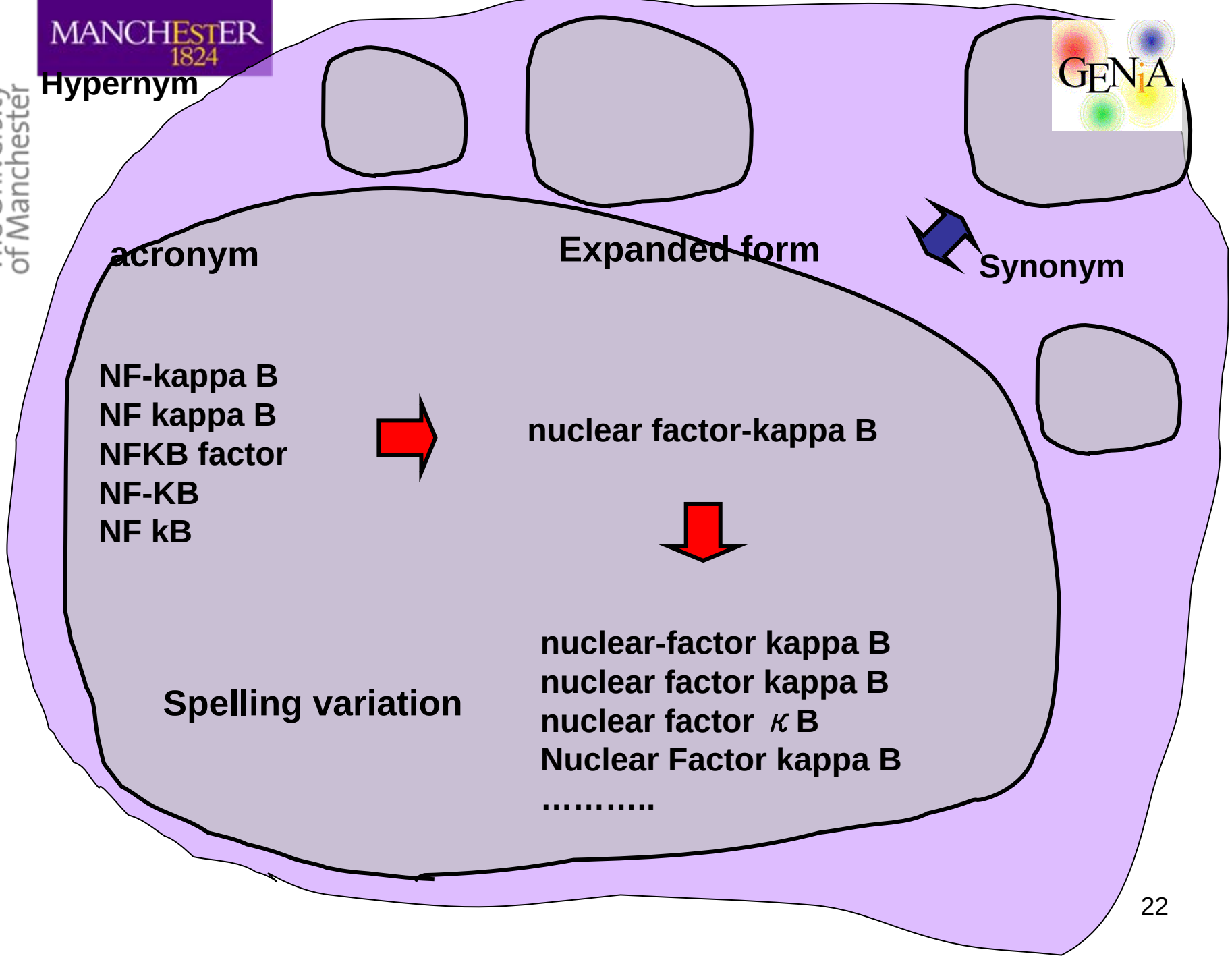
Language Domain

Linguistic expressions

Knowledge Domain

Concepts and Relationships among Them

Motivated Independently of language

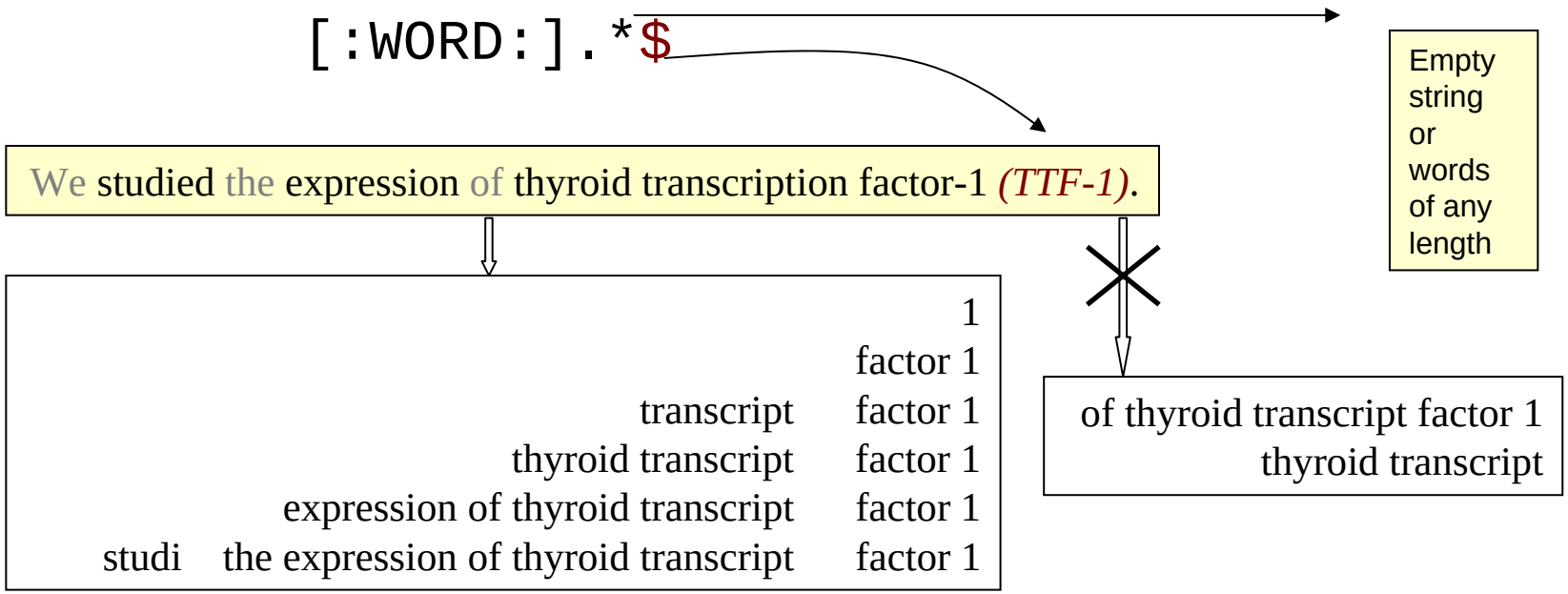

Hypernym
acronym
Expanded form
Synonym
NF-kappa B
NF kappa B
NFKB factor
NF-KB
NF kB
nuclear factor-kappa B
Spelling variation
nuclear-factor kappa B
nuclear factor kappa B
nuclear factor κ B
Nuclear Factor kappa B
……

Acronym Dictionary

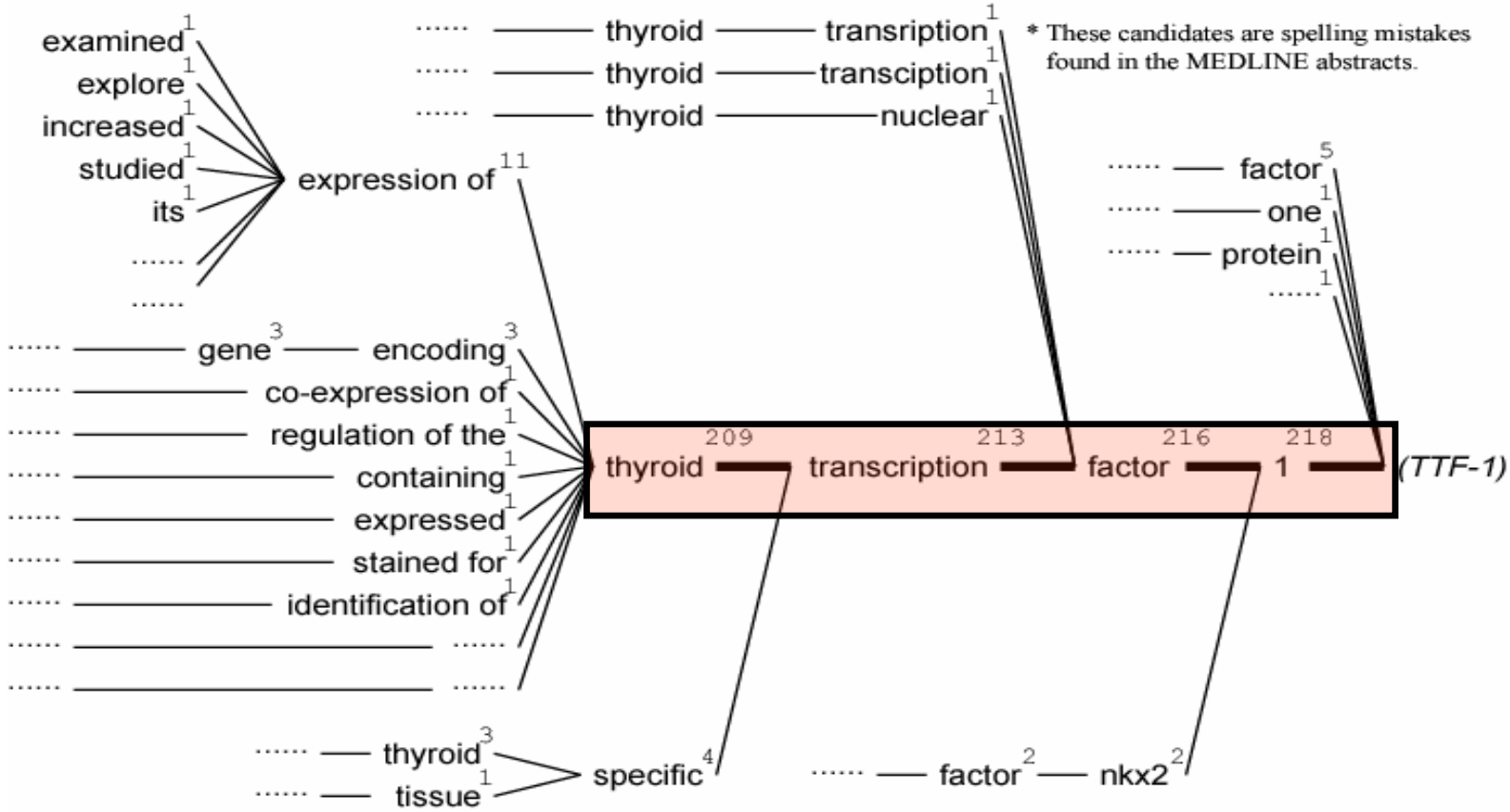
- N.Okazaki, S.Ananiadou (NaCTeM)
  - Statistics-based Acronym recognizer (not rule-based)

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



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

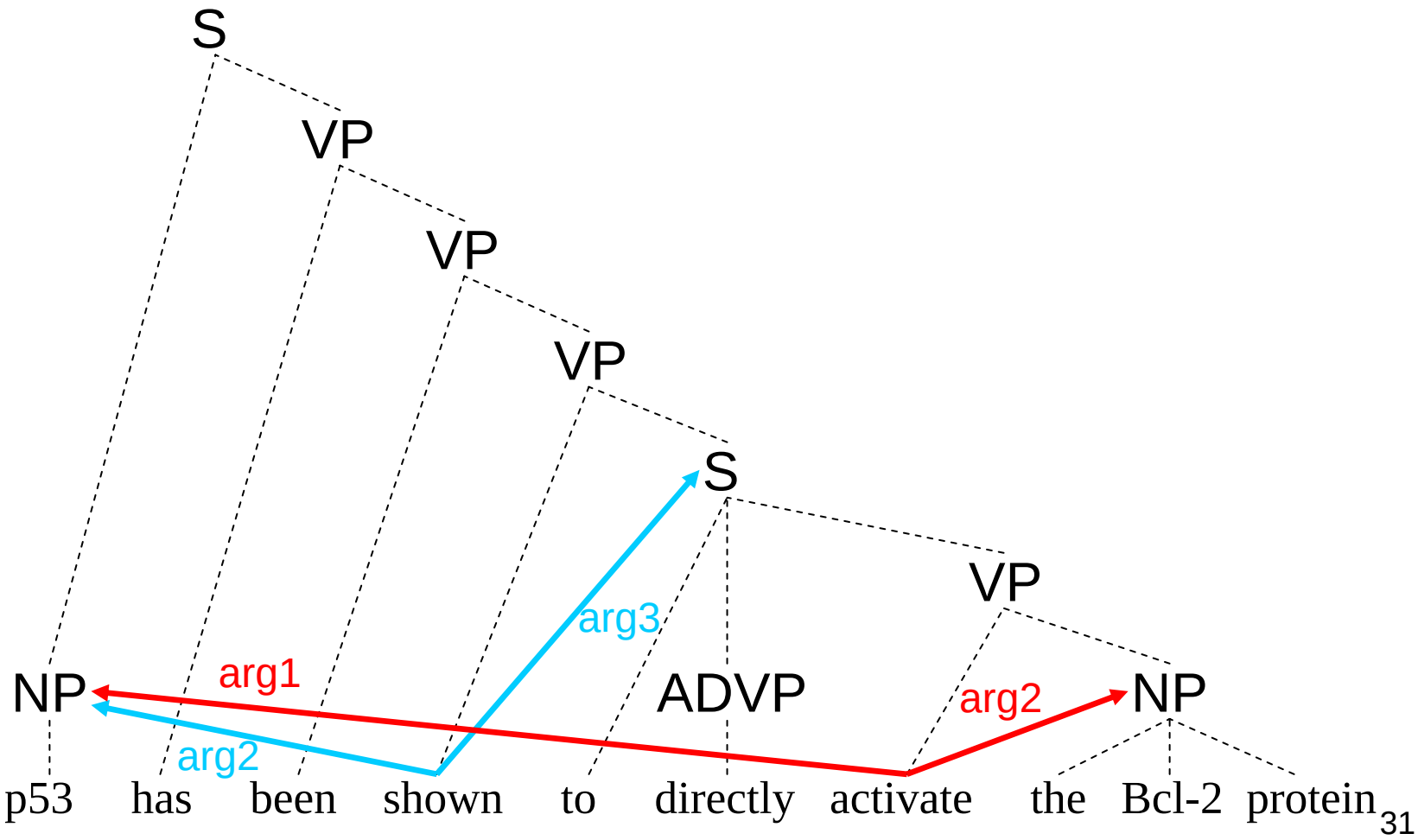
1.000	Human immunodeficiency virus type 1	4
0.400	HIV type 1	52
0.213	HLA - Human immunodeficiency virus type 1	0
0.114	Human immunodeficiency virus type-1	5
0.105	Human immuno virus type 1	0
0.085	HLA - HIV type 1	0
0.077	Human immunodeficiency virus 1	3
0.074	Human immunodeficiencydeficiency virus 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 Precison	39.2%	31.8%
Processing Time	0.68sec	1.00sec

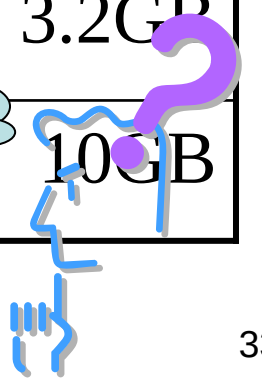
Scalability of TM Tools

Target Corpus: MEDLINE corpus

The number of papers	14,792,890
The number of abstracts	7,434,879
The number of ser	,480
The number of y	,650
Comp	3.2GB
Un	10GB

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

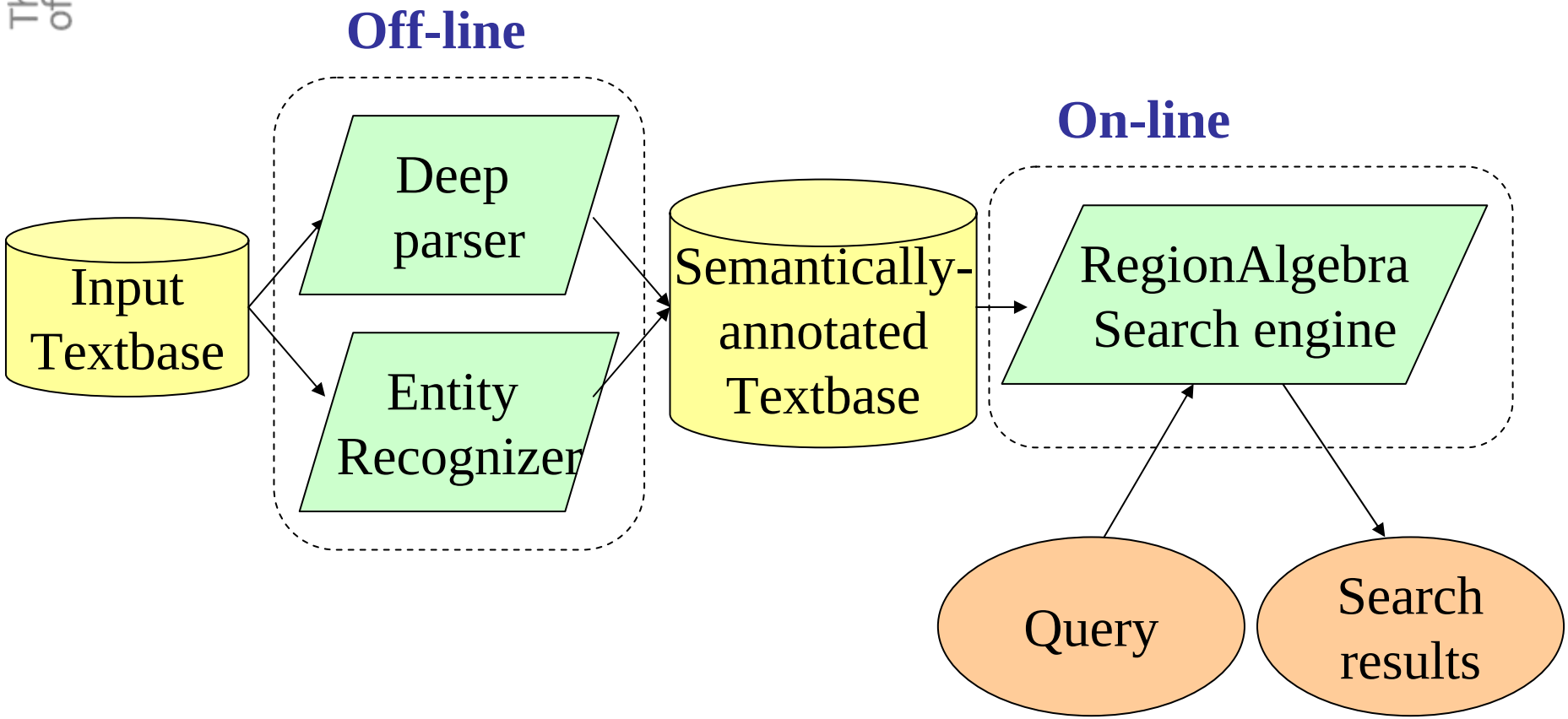
70 million seconds, that is, about 2 years



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.

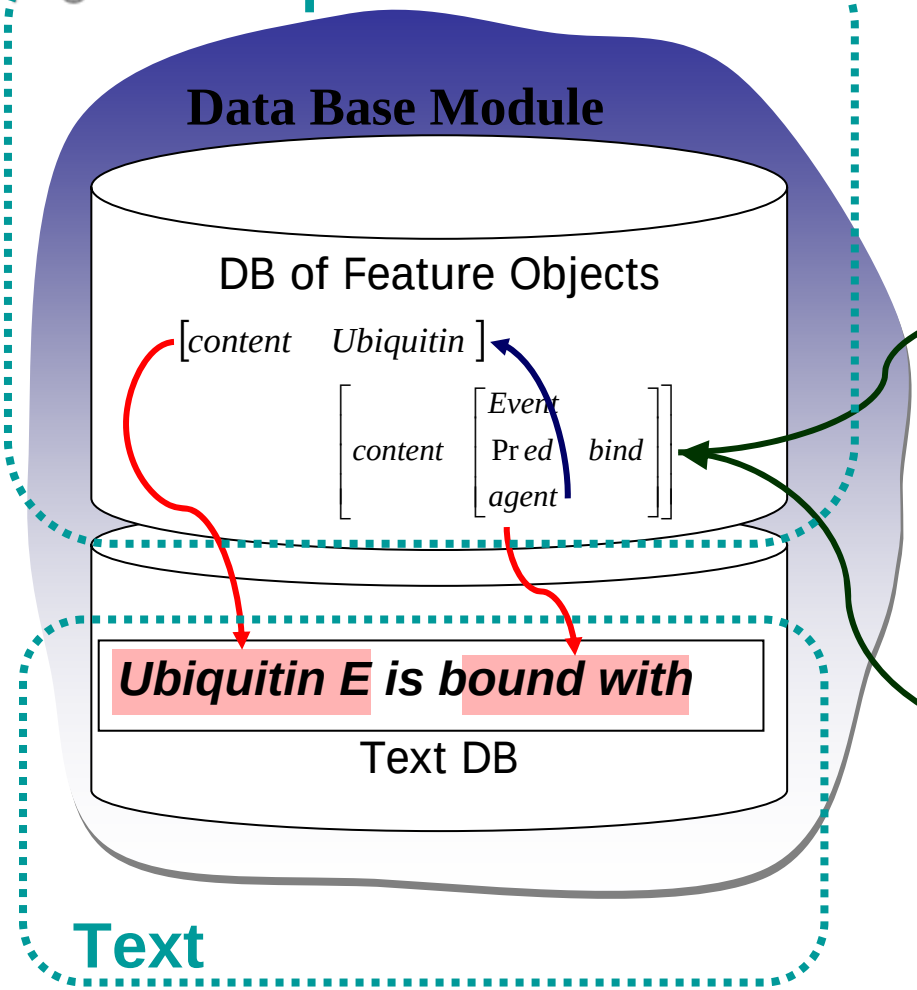
Medie system overview



Text Archive with Feature Obejcts

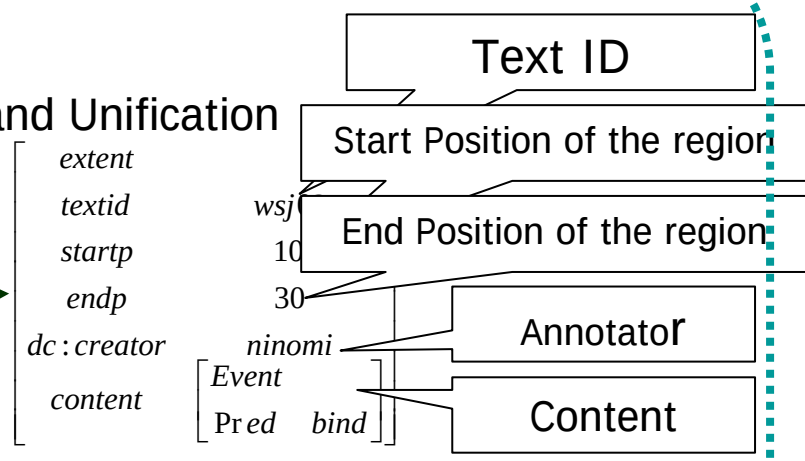
Managing texts, data representation and their semantics

Data representation

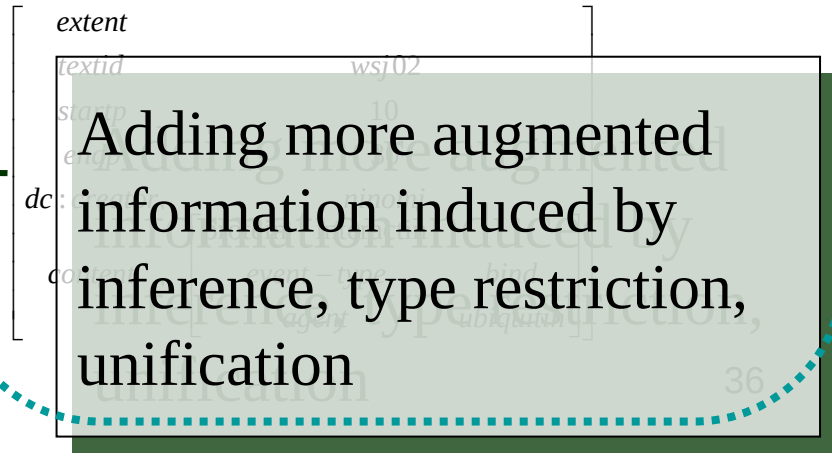


Semantics

Copy and Unification



Specialization by unification



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

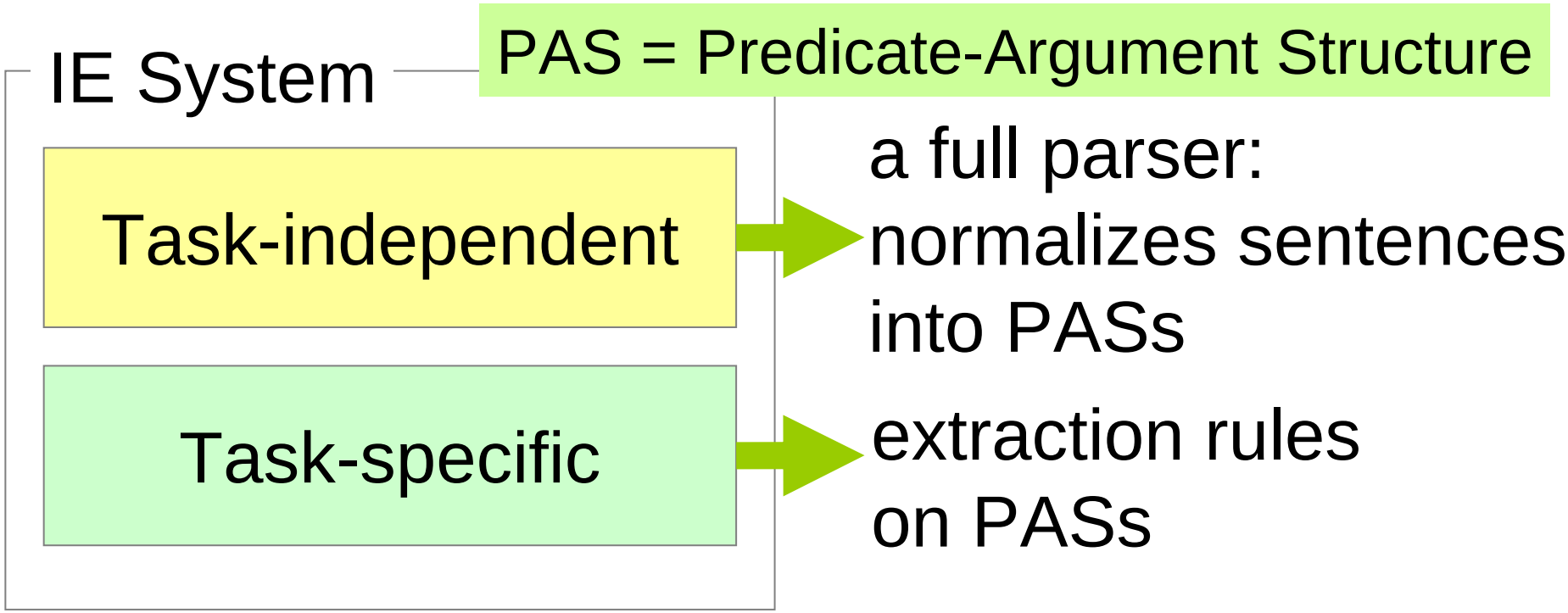
Text

Demos

- MEDIE
- Info-PubMed
- TerMine

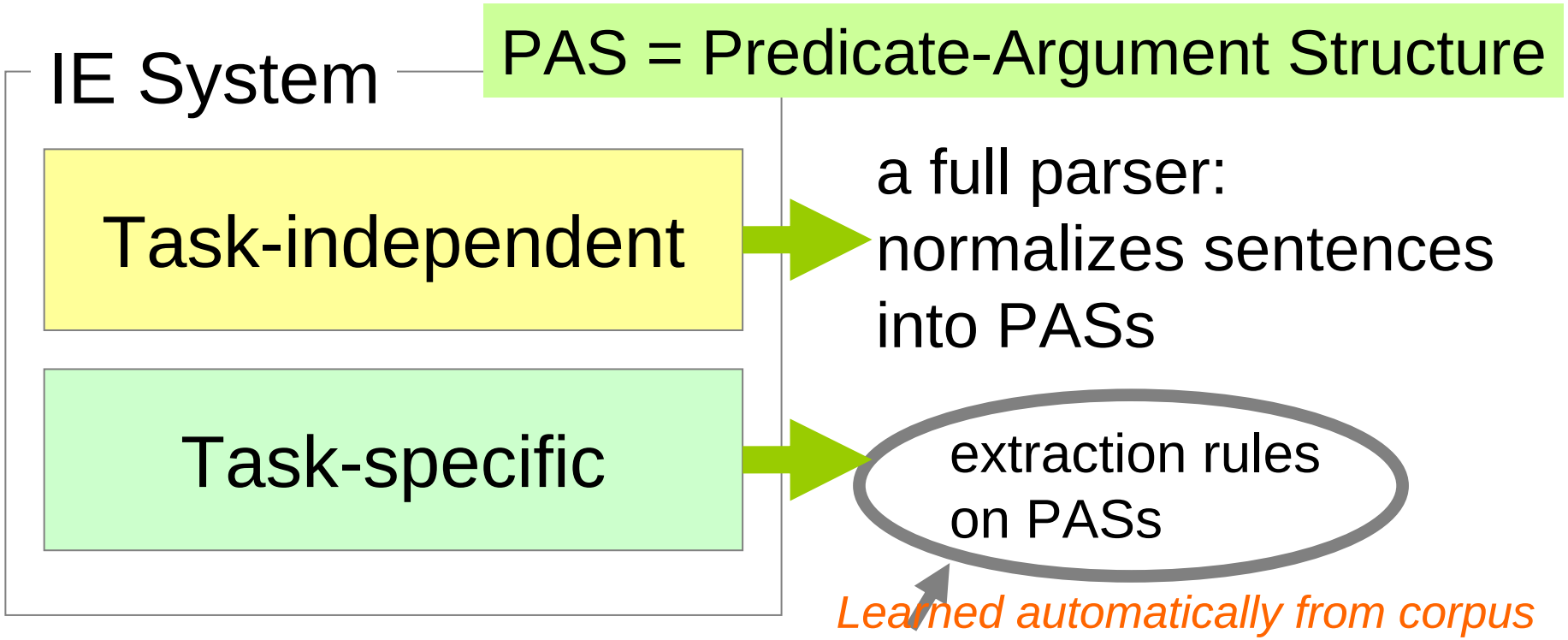
Our Policy for IE

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



Our Policy for IE

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



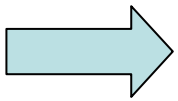
Advantages of Full Parsing

- 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*

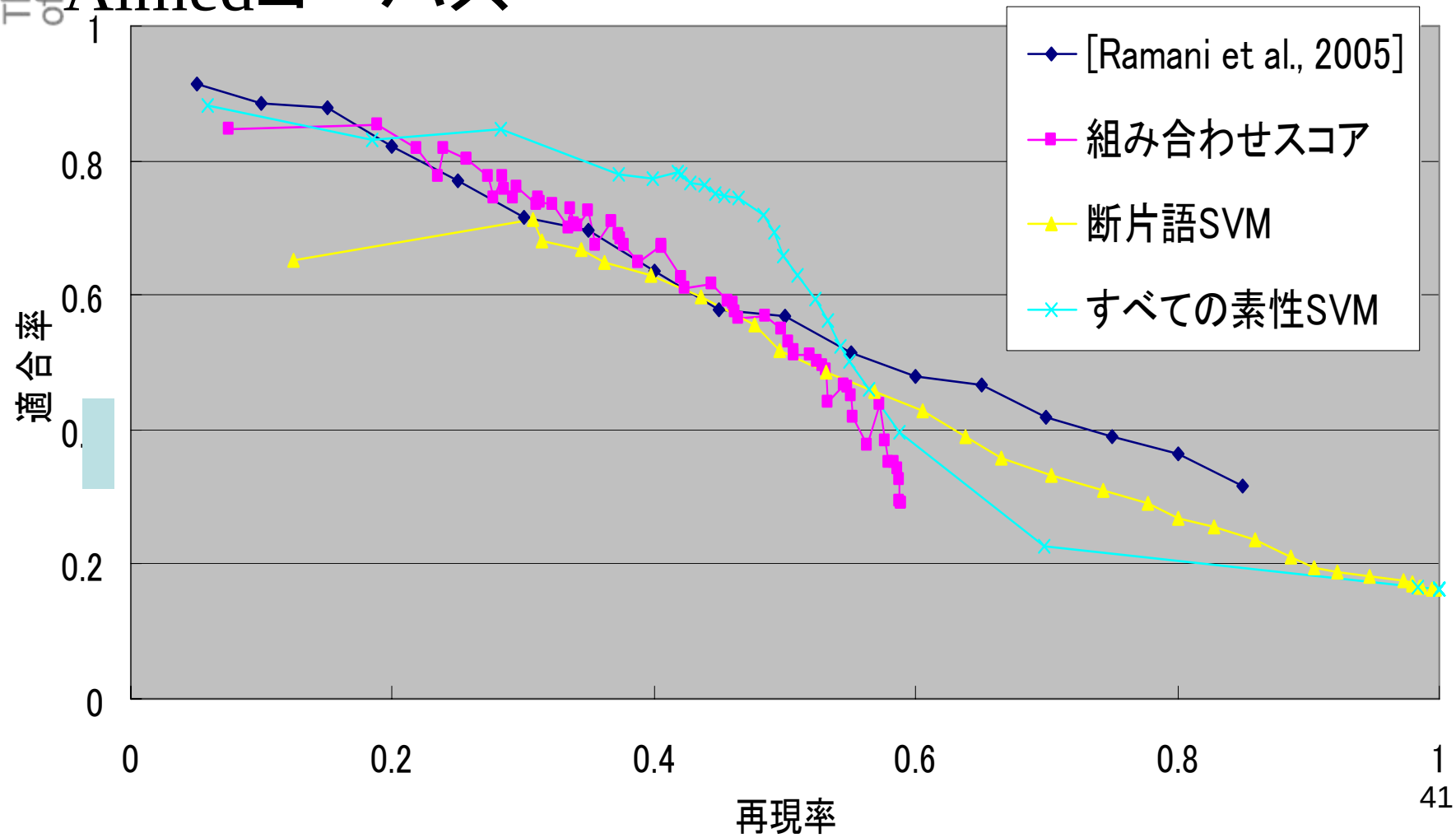
We can construct more general extraction rules.
Less extraction rules



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)

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

  - 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

Training corpus	WSJ	GENIA
WSJ	97.0	84.3
GENIA	75.2	98.1
WSJ+GENIA	96.9	98.1

No degradation of the tagger trained by the mixed corpus

(Ref.) TnT tagger

Training corpus	WSJ	GENIA
WSJ	96.7	84.3
GENIA	80.1	97.9
WSJ+GENIA	96.5	97.5

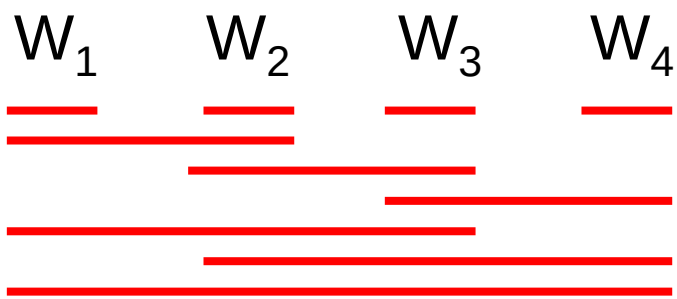
Some degradations (0.2 ~ 0.4) were observed, compared with the taggers trained by "pure" corpora

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
PROTEIN DNA
recruits another B cell-specific coactivator to increase transcriptional activity of class II promoters in B cells."
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

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

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

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

$w_b = X$ $\quad$ $w_{e+2} = X$

$w_e = X$

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

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
Our method	**71.5**	**70.2**	**70.8**
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 Precison	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 combiined by Reference distribution

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

Weights

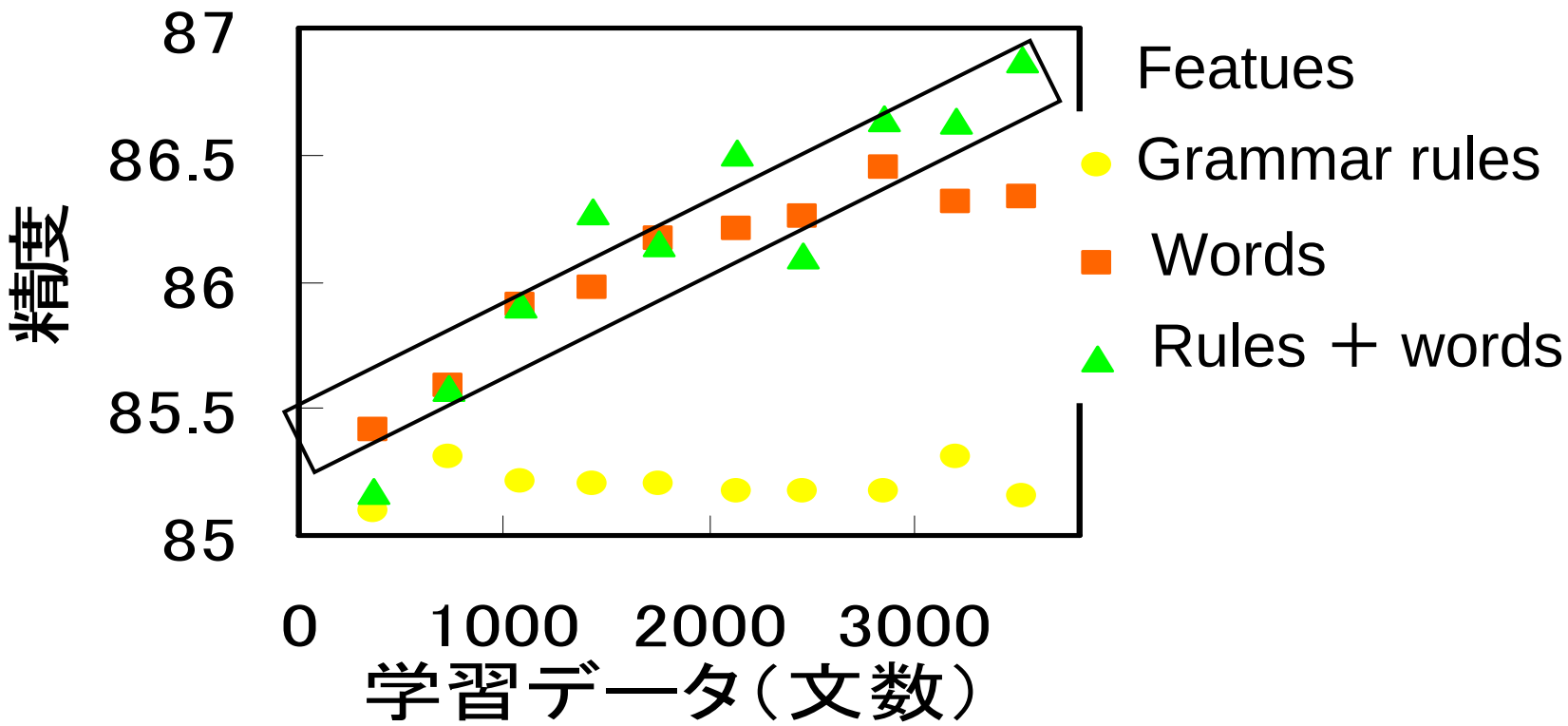
features

reference distribution

- The new model maximizes the model as a whole

Corpus Size and Precision

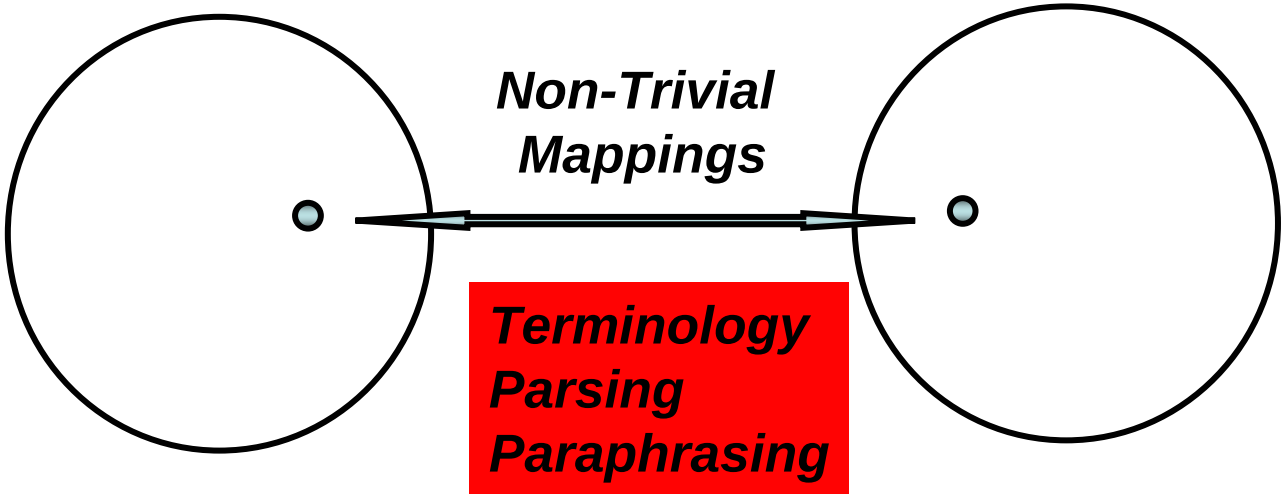
- Domain Adaptation Experiment using GENIA
- Less Learning Time, The Precision improves rapidly



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

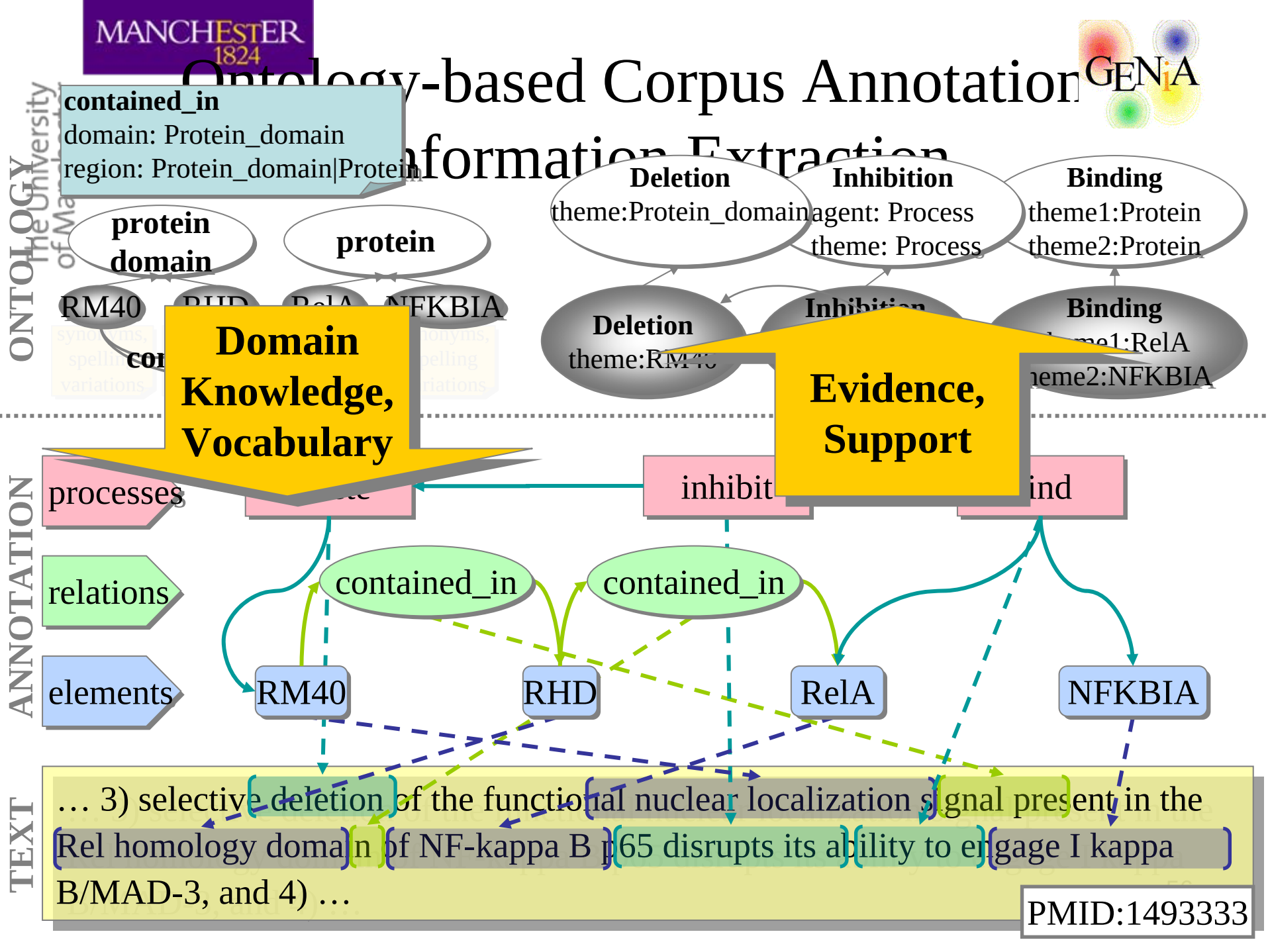
Linguistic expressions

Knowledge Domain

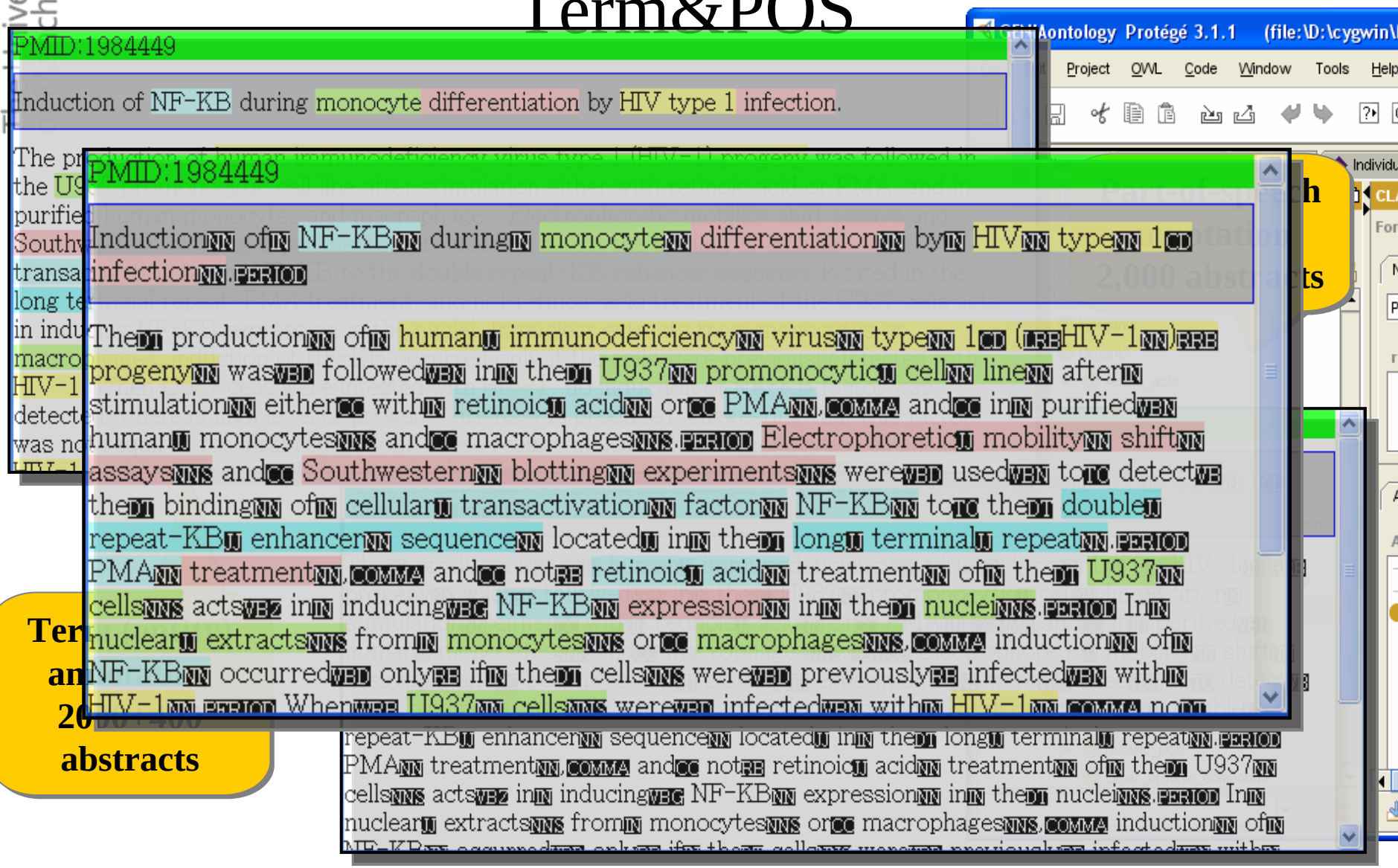
Concepts and Relationships among Them

Motivated Independently of language

Ontology-based Corpus Annotation for Information Extraction

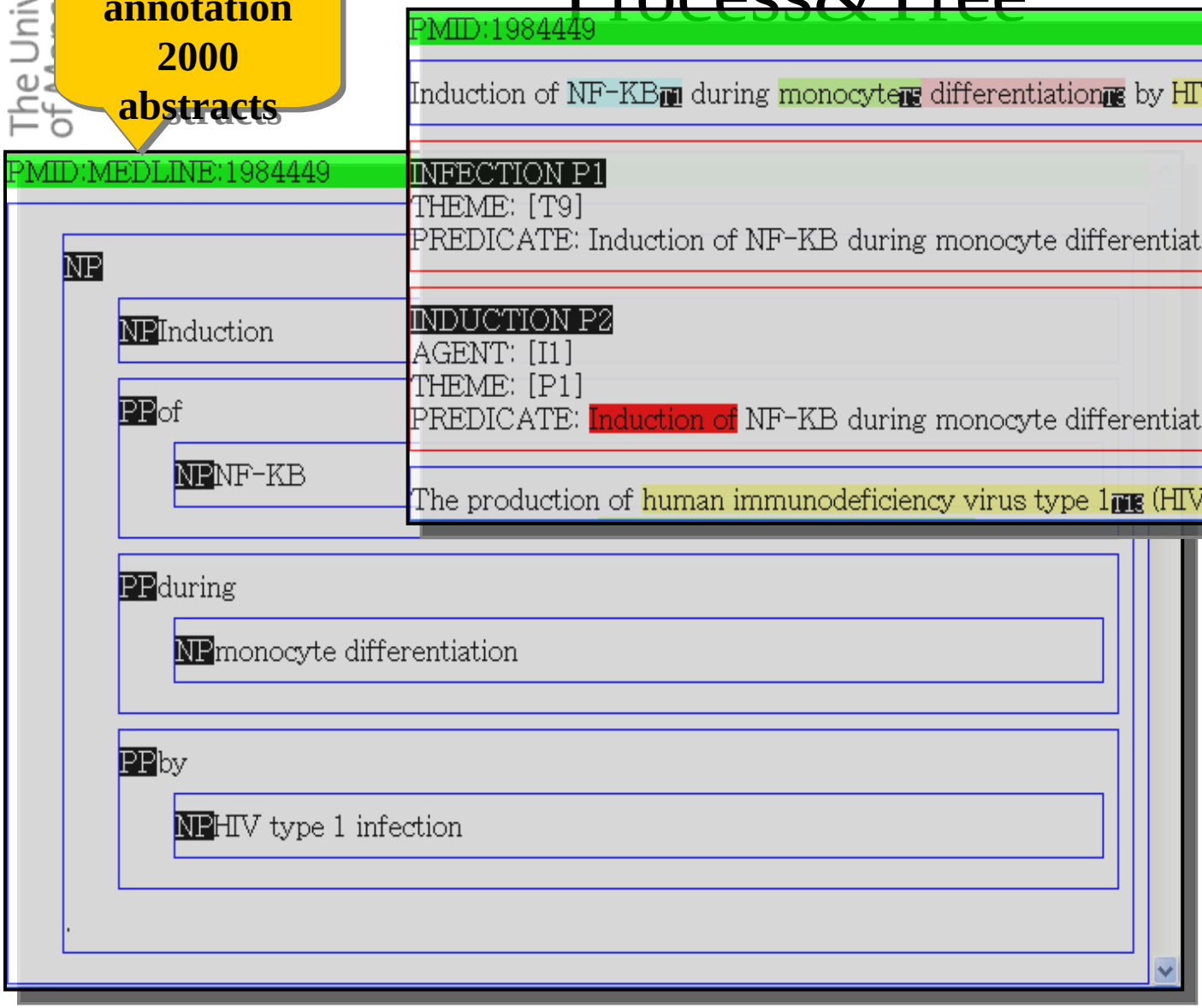


Annotation of GENIA corpus – Term&POS



Annotation of GENIA corpus – Process&Tree

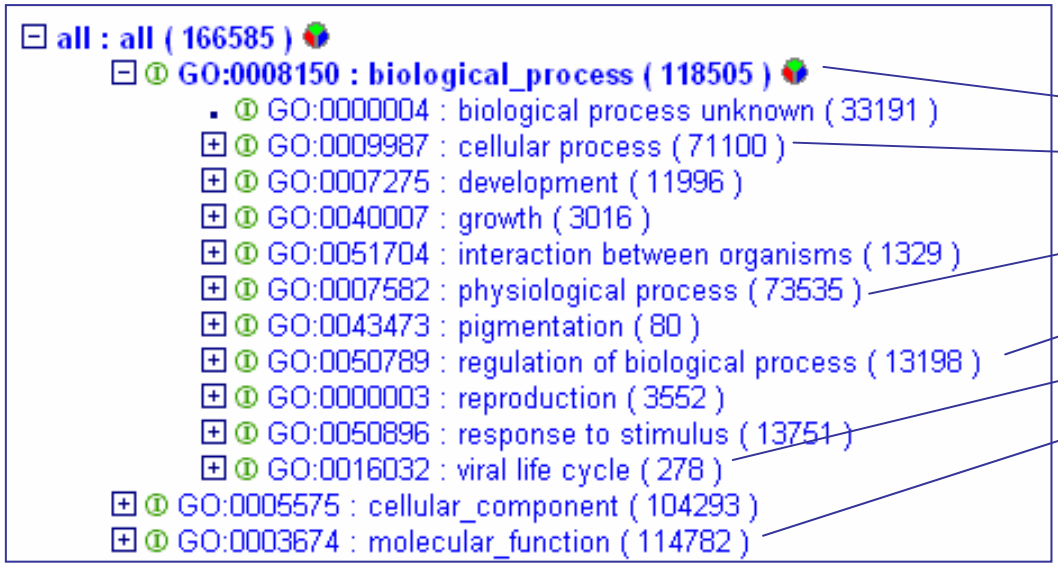
Tree annotation 2000 abstracts



Process annotation 500 abstracts (preliminary)

GENIA event ontology

Gene ontology



GENIA ontology

+Class (35)
- +Biological_process (31)
- +Cellular_process (2)
- +Physiological_process (21)
- +Regulation (2)
- +Viral_life_cycle (2)
- +Molcular_function (2)
- +Artificial_process
- +Correlation

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

Text Mining

Text as a bag of words

Words as surface strings

NLP-based TM

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