Fine-Grained Information Access and MT for Patents

Junichi TSUJII
University of Tokyo,
University of Manchester
UK National Centre for Text Mining

Plan of the Talk

- Backgrounds
  - Shifts in Society
  - Semantics-based information access to text and MT
- NLP and Semantics-based information access
  - Deep parsing and entity/event-based access
  - Efficient deep parsers
- MT with Deep Parsing
  - Structural Differences: Predicate-Argument Structures
  - Lexical Differences: Terms and Ontologies
- Concluding remarks and future work

Plan of the Talk

- Backgrounds
  - Shifts in Society
  - Semantics-based information access to text and MT
- NLP and Semantics-based information access
  - Deep parsing and entity/event-based access
  - Efficient deep parsers
- MT with Deep Parsing
  - Structural Differences: Predicate-Argument Structures
  - Lexical Differences: Terms and Ontologies
- Concluding remarks and future work

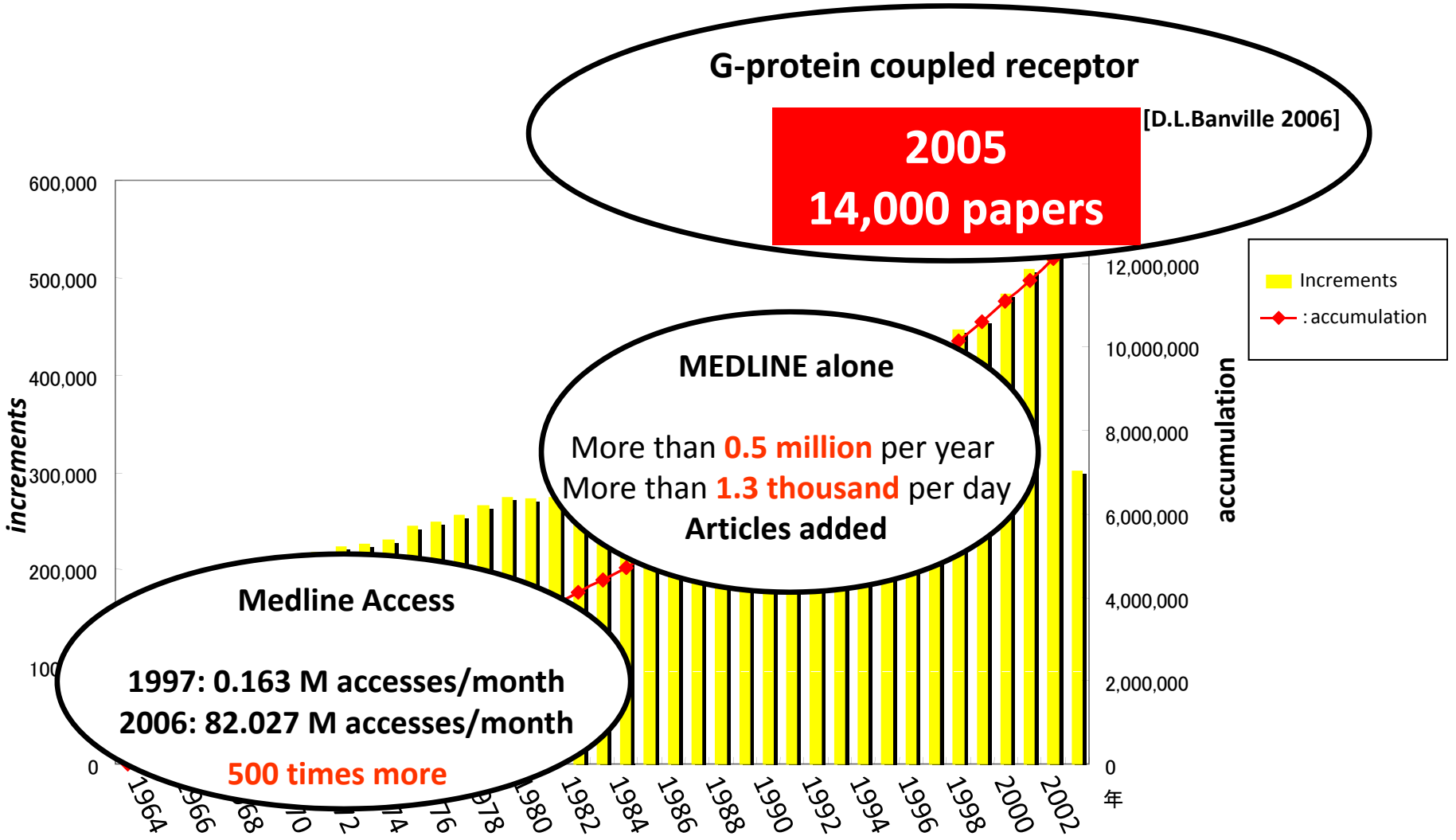
MT for Patents among Asian Languages

- Shift in society
  - From Labor-based Industrial Society to Knowledge-based Industrial Society
- Shift in the world
  - Globalization
  - From the Economies in the Atlantic-rim to the Economies in the Pacific-rim
  - Translation among (Asian Languages + English)

- Difficulties in MT
  - Translation among European languages to Translation among (Asian Languages + English)
    - Structural and Lexical differences
    - Sentence-oriented languages and Context-oriented languages
- **SMT with great enhancement of semantics-based processing**

Plan of the Talk

- Backgrounds
  - Shifts in Society
  - Semantics-based information access to text and MT
- NLP and Semantics-based information access
  - Deep parsing and entity/event-based access
  - Efficient deep parsers
- MT with Deep Parsing
  - Structural Differences: Predicate-Argument Structures
  - Lexical Differences: Terms and Ontologies
- Concluding remarks and future work



Increase in Medline

NaCTeM

www.nactem.ac.uk

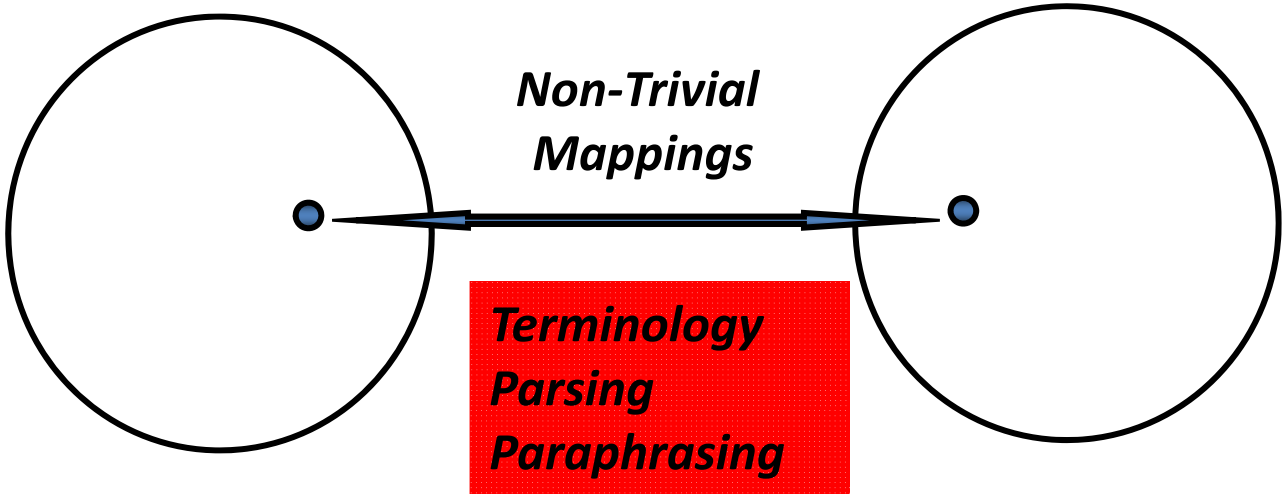
- First such centre in the world
- Funding: JISC, BBSRC, EPSRC
- Consortium investment
    - Chair in TM (Prof. J. Tsujii, Univ. Tokyo)
- Location: Manchester Interdisciplinary Biocentre (MIB) www.mib.ac.uk funded by the Wellcome Trust
- Initial focus: biomedical academic community
- Extend services to industry
- Extend focus to other domains (social sciences)

Consortium

- Universities of Manchester, Liverpool
- Service activity run by MIMAS (National Centre for Dataset Services), within MC (Manchester Computing)
- Self-funded partners
  - San Diego Supercomputing Center
  - University of California, Berkeley
  - University of Geneva
  - University of Tokyo
- Strong industrial & academic support
  - IBM, AZ, EBI, Wellcome Trust, Sanger Institute, Unilever, NowGEN, MerseyBio, …

Semantics-based fine-grained information access

From surface diversities and ambiguities to conceptual invariants



Language Domain

Linguistic expressions

Knowledge Domain

Concepts and Relationships among Them

Motivated Independently of language

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

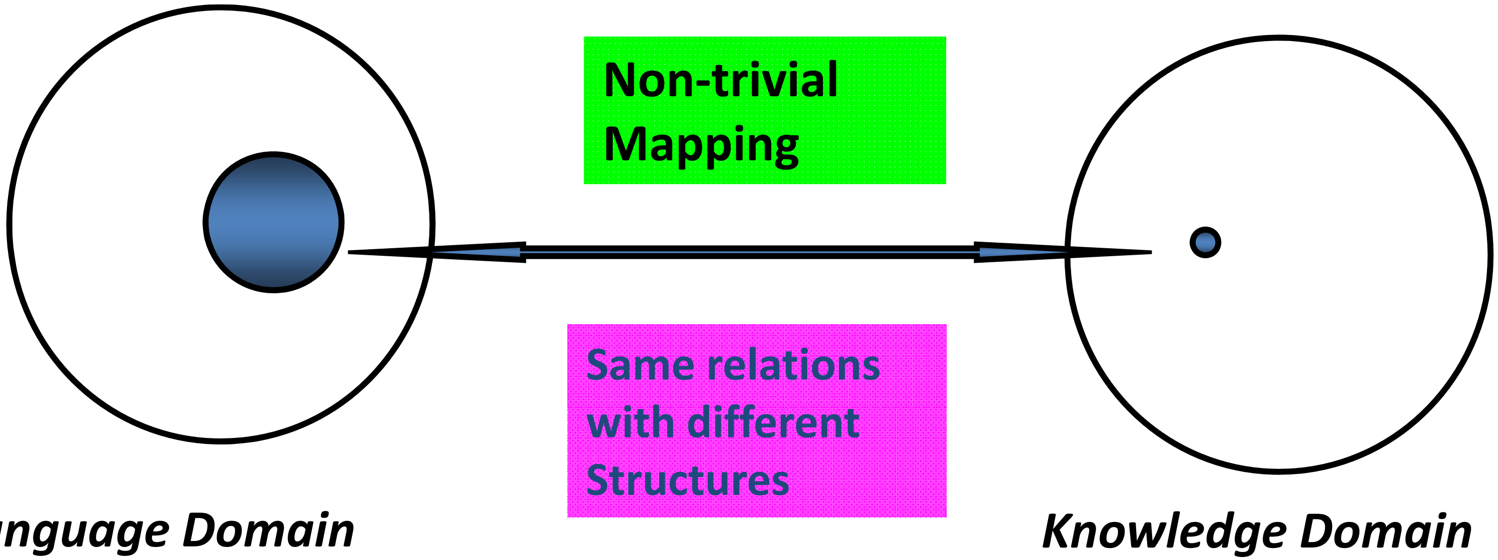
Transcription initiation by the sigma(54)-RNA polymerase holoenzyme requires **an enhancer-binding protein** that is thought to contact sigma(54) to **activate** transcription.

Full-strength Staufen protein lacking this insertion is able to assocaite with osker mRNA and **activate** its translation, but fails to ...

PHO2 protein activates transcription of PHO5 gene.

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

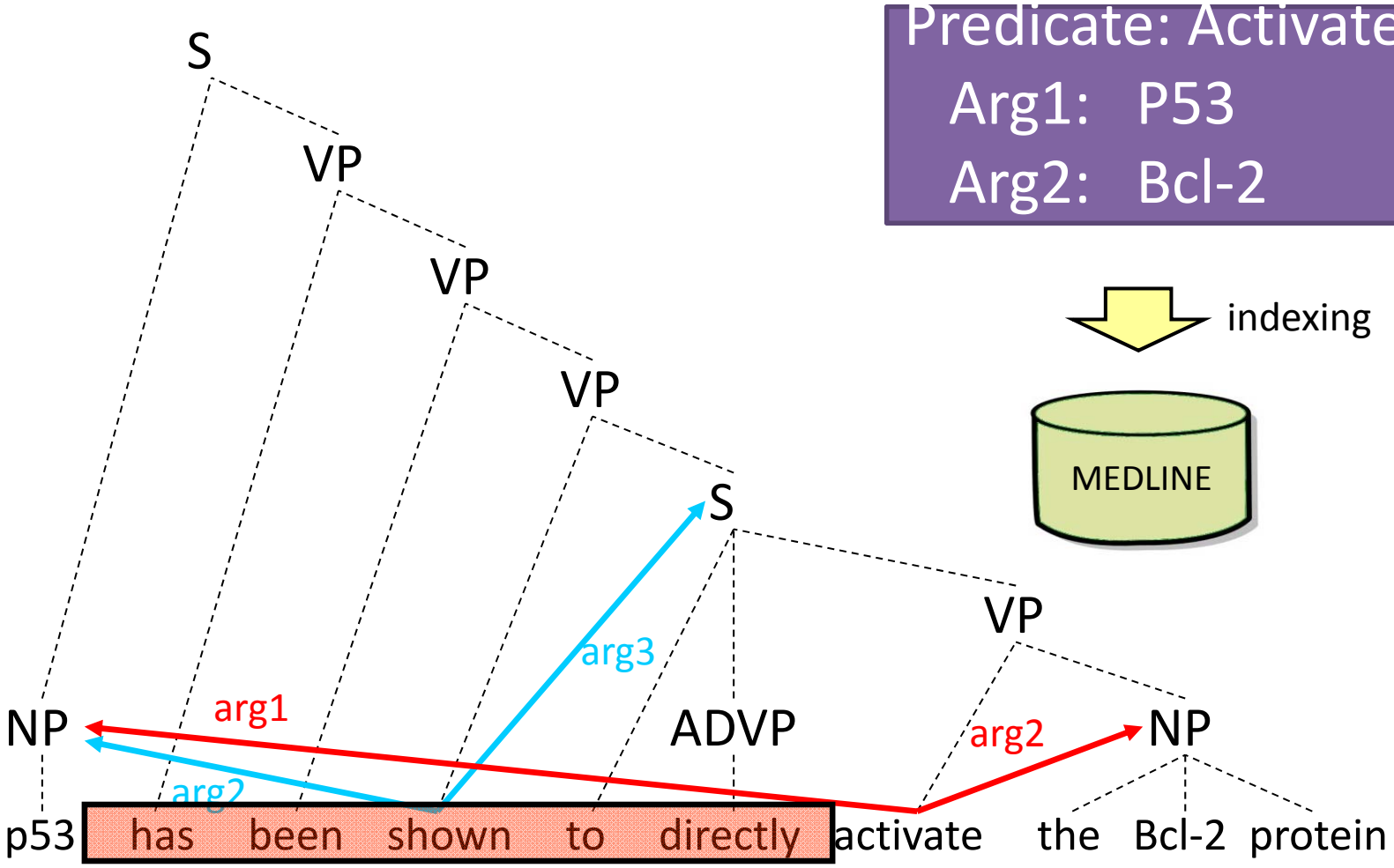
Retrieval using Regional Algebra

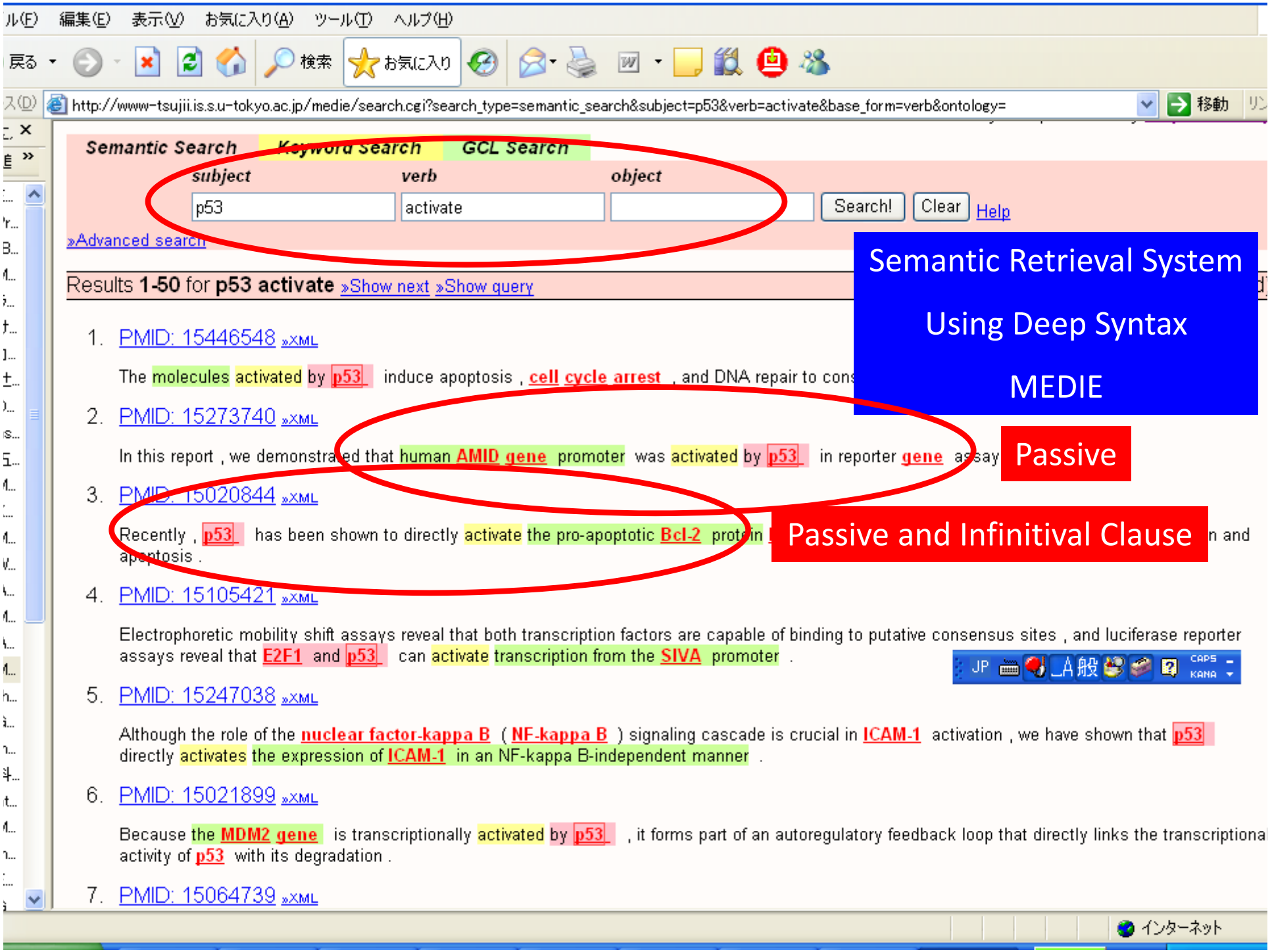


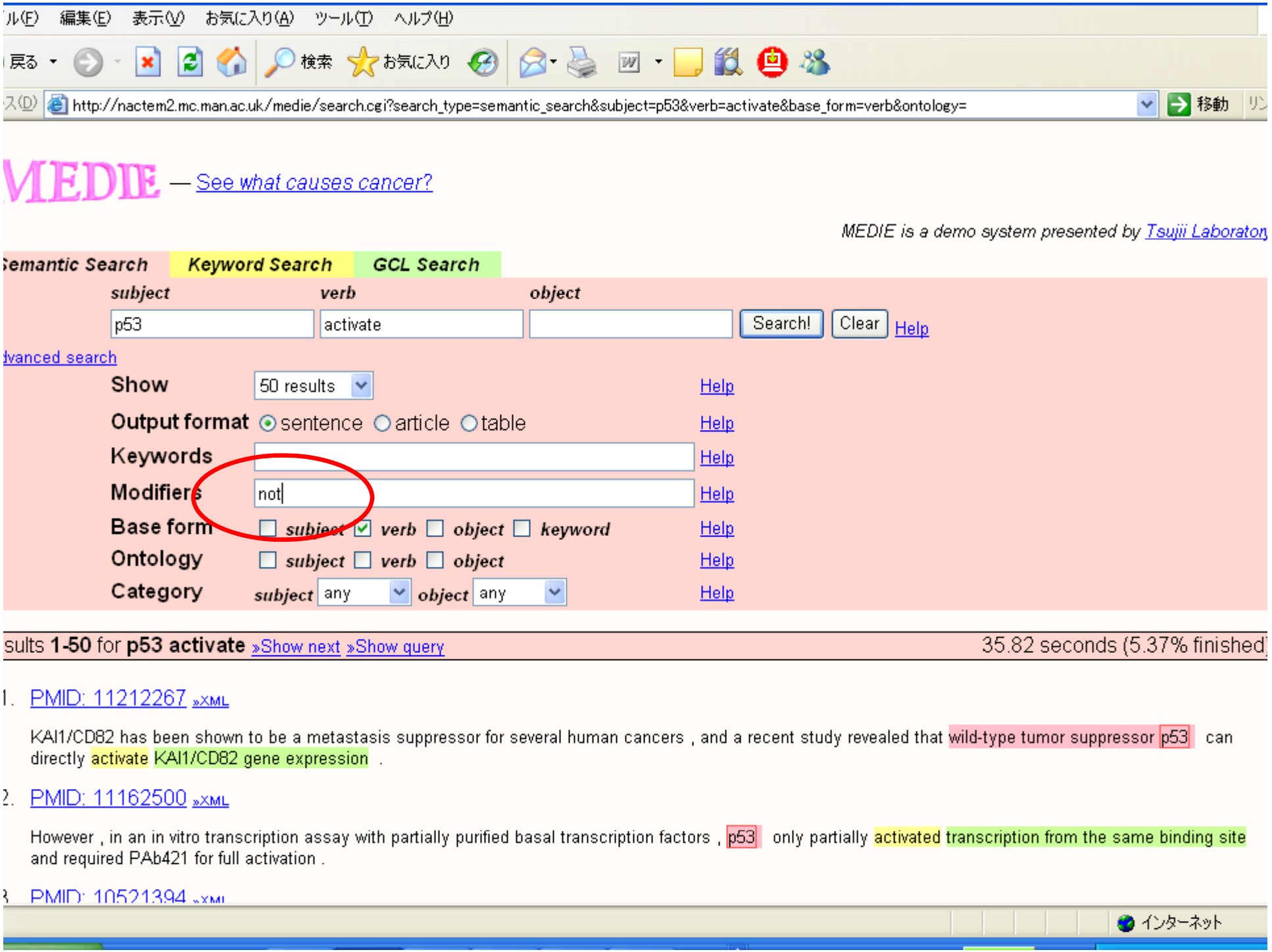
Independently motivated of Language

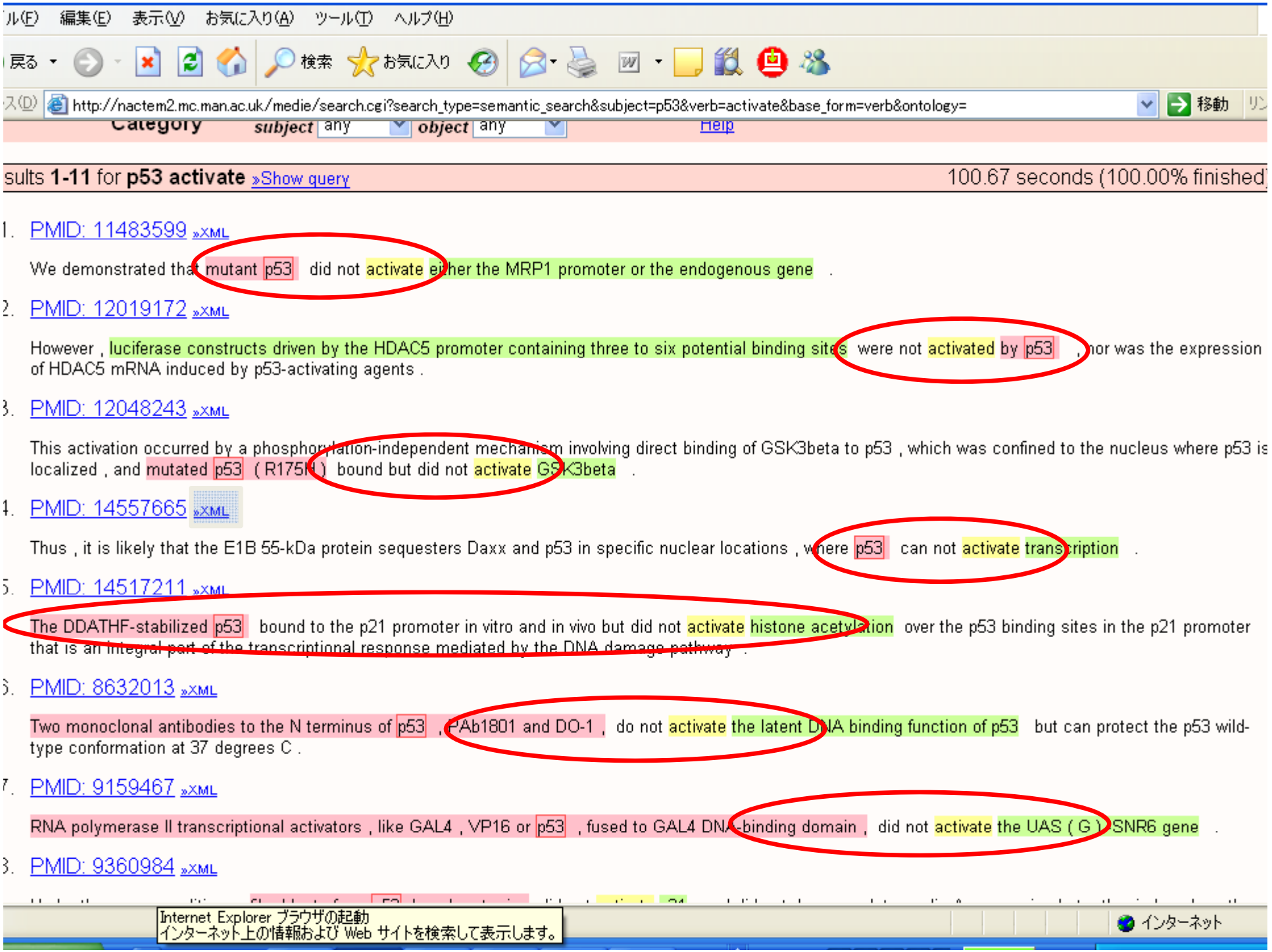
Deep Parsing

Parser based on Probabilistic HPSG (Enju)



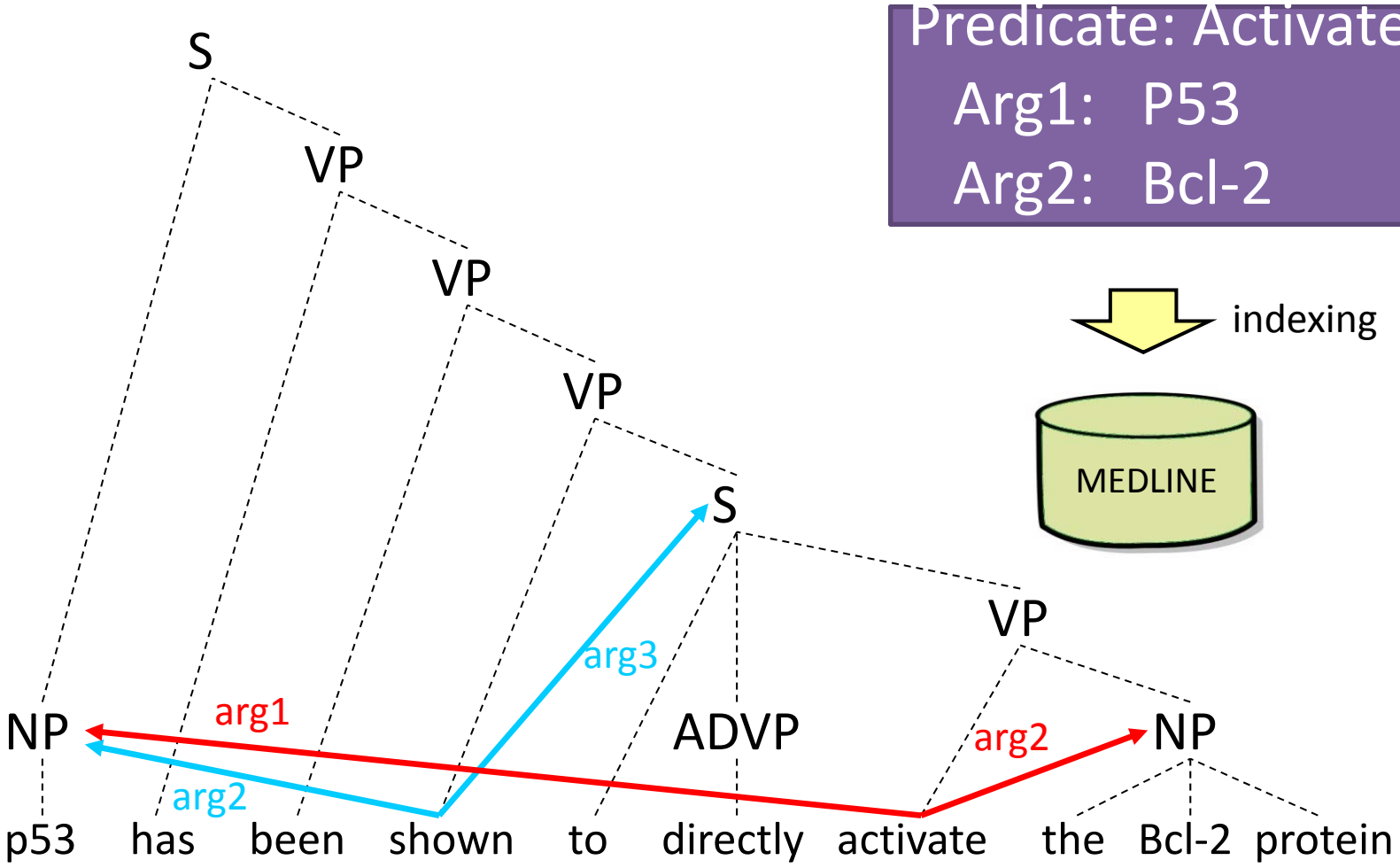

Semantic Search
Keyword Search
GCL Search
subject
verb
object
p53
activate
Search!
Clear
Help
»Advanced search
Results 1-50 for p53 activate »Show next »Show query
1. PMID: 15446548 »XML
The molecules activated by p53 induce apoptosis , cell cycle arrest , and DNA repair to cons
2. PMID: 15273740 »XML
In this report , we demonstrated that human AMID gene promoter was activated by p53 in reporter gene assay
3. PMID: 15020844 »XML
Recently , p53 has been shown to directly activate the pro-apoptotic Bcl-2 prot in
n and
apoptosis .
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
Semantic Retrieval System
Using Deep Syntax
MEDIE
Passive
Passive and Infinitival Clause
インターネット


ル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
http://nactem2.mc.man.ac.uk/medie/search.cgi?search_type=semantic_search&subject=p53&verb=activate&base_form=verb&ontology=
移動
MEDIE — See what causes cancer?
MEDIE is a demo system presented by Tsujii Laboraton
emantic Search
Keyword Search
GCL Search
subject
verb
object
p53
activate
Search!
Clear
Help
dvanced search
Show
50 results
Help
Output format
sentence
article
table
Help
Keywords
Help
Modifiers
not
Help
Base form
subject
verb
object
keyword
Help
Ontology
subject
verb
object
Help
Category
subject
any
object
any
Help
sults 1-50 for p53 activate »Show next »Show query
35.82 seconds (5.37% finished)
1. PMID: 11212267 »XML
KAI1/CD82 has been shown to be a metastasis suppressor for several human cancers , and a recent study revealed that wild-type tumor suppressor p53 can directly activate KAI1/CD82 gene expression .
2. PMID: 11162500 »XML
However , in an in vitro transcription assay with partially purified basal transcription factors , p53 only partially activated transcription from the same binding site and required PAb421 for full activation .
PMID: 10521394 »XML
インターネット


ル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
http://nactem2.mc.man.ac.uk/medie/search.cgi?search_type=semantic_search&subject=p53&verb=activate&base_form=verb&ontology=
移動
Category
subject any
object any
Help
sults 1-11 for p53 activate »Show query
100.67 seconds (100.00% finished)
1. PMID: 11483599 »XML
We demonstrated that mutant p53 did not activate either the MRP1 promoter or the endogenous gene .
2. PMID: 12019172 »XML
However , luciferase constructs driven by the HDAC5 promoter containing three to six potential binding sites were not activated by p53 , nor was the expression of HDAC5 mRNA induced by p53-activating agents .
3. PMID: 12048243 »XML
This activation occurred by a phosphorylation-independent mechanism involving direct binding of GSK3beta to p53 , which was confined to the nucleus where p53 is localized , and mutated p53 (R175H) bound but did not activate GSK3beta .
4. PMID: 14557665 »XML
Thus , it is likely that the E1B 55-kDa protein sequesters Daxx and p53 in specific nuclear locations , where p53 can not activate transcription .
5. PMID: 14517211 »XML
The DDATHF-stabilized p53 bound to the p21 promoter in vitro and in vivo but did not activate histone acetylation over the p53 binding sites in the p21 promoter that is an integral part of the transcriptional response mediated by the DNA damage pathway .
6. PMID: 8632013 »XML
Two monoclonal antibodies to the N terminus of p53 , PAb1801 and DO-1 , do not activate the latent DNA binding function of p53 but can protect the p53 wild-type conformation at 37 degrees C .
7. PMID: 9159467 »XML
RNA polymerase II transcriptional activators , like GAL4 , VP16 or p53 , fused to GAL4 DNA-binding domain , did not activate the UAS (G) -SNR6 gene .
8. PMID: 9360984 »XML
Internet Explorer ブラウザの起動
インターネット上の情報および Web サイトを検索して表示します。
インターネット

Deep Parsing

Parser based on Probabilistic HPSG (Enju)



Results on PTB-WSJ

Parser	grammar	Accuracy	Speed
MST parser	dependency	90.02% (LAS)	4.5 snt/sec
Sagae's parser	dependency	89.01% (LAS)	21.6 snt/sec
Berkeley parser	CFG	89.27% (LF_1)	4.7 snt/sec
Charniak's parser	CFG	89.55% (LF_1)	2.2 snt/sec
Charniak's parser reranker	CFG	91.40 % (LF_1)	1.9 snt/sec
Enju parser	HPSG	88.87% (PAS-LF_1)	2.7 snt/sec
Mogura parser	HPSG	88.07% (PAS-LF_1)	22.8 snt/sec

Plan of the Talk

- Backgrounds
  - Shifts in Society
  - Semantics-based information access to text and MT
- NLP and Semantics-based information access
  - Deep parsing and entity/event-based access
  - Efficient deep parsers
- MT with Deep Parsing
  - Structural Differences: Predicate-Argument Structures
  - Lexical Differences: Terms and Ontologies
- Concluding remarks and future work

Fine-grained tree-to-string translation rule extraction

(presented in ACL 2010)

Xianchao Wu (吴 先超)

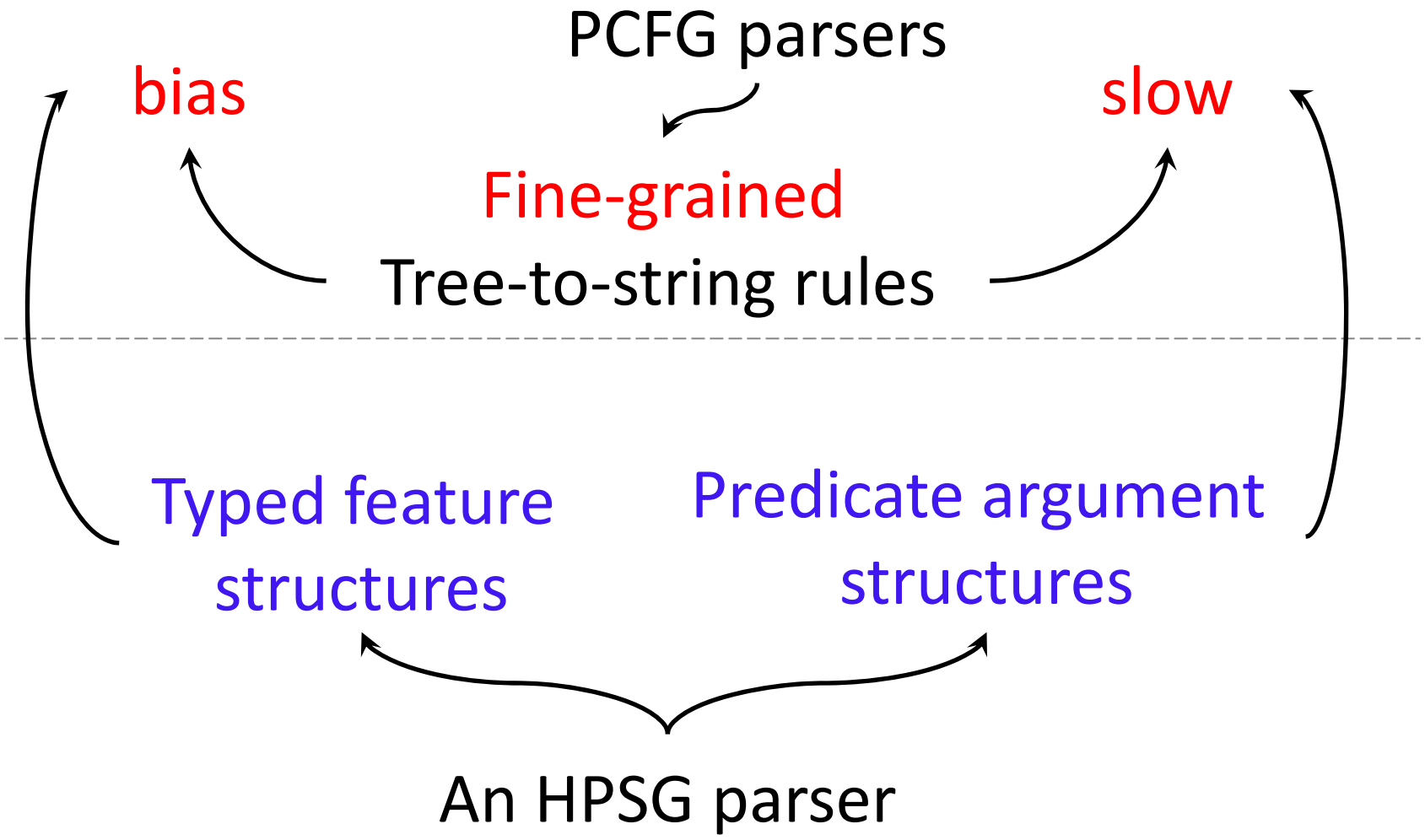
Takuya Matsuzaki, Jun'ichi Tsujii

The University of Tokyo

2010.07.23

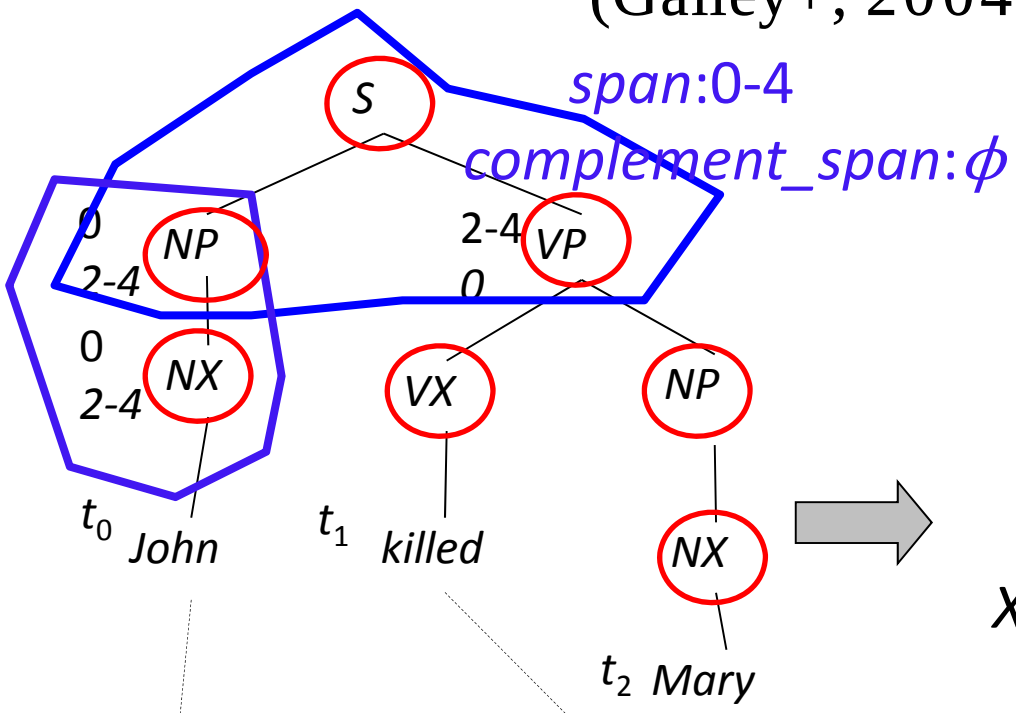
東京大学 THE UNIVERSITY OF TOKYO

Outline



Minimal rule extraction : GHKM

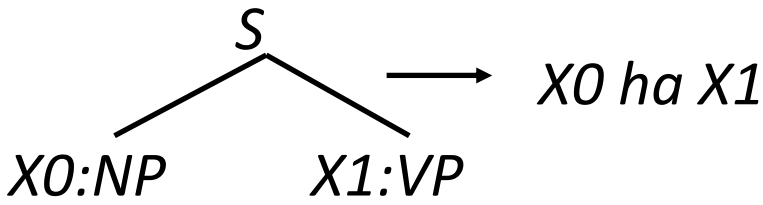
(Galley+, 2004)

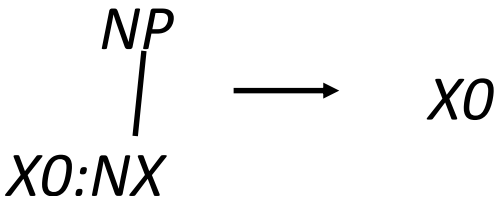


NO overlapping

↓

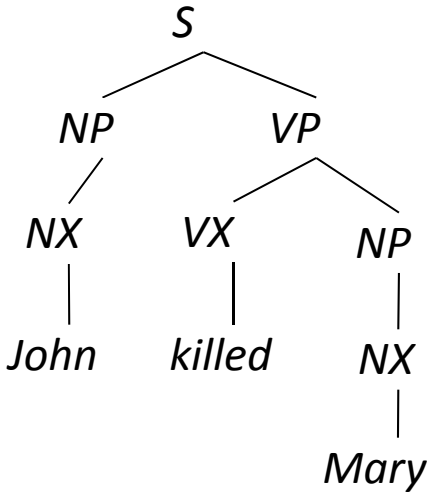
Admissible nodes



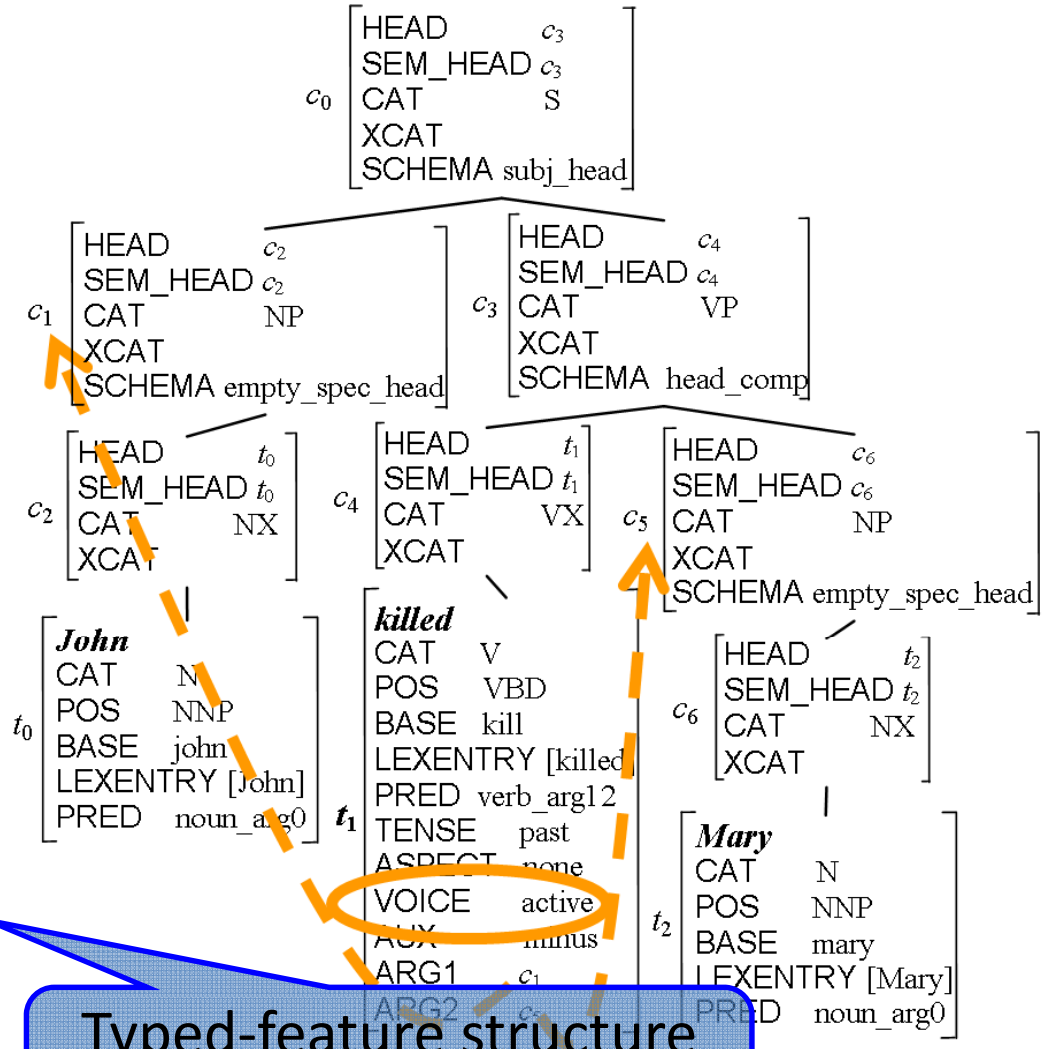


東京大学 THE UNIVERSITY OF TOKYO

HPSG for syntax-based SMT



nodes	features
Non-terminal	CAT, XCAT, SCHEMA, HEAD, SEM_HEAD
terminal	CAT, POS, BASE, TENSE, ASPECT, VOICE, AUX, LEXENTRY, PRED, ARG<*n*>

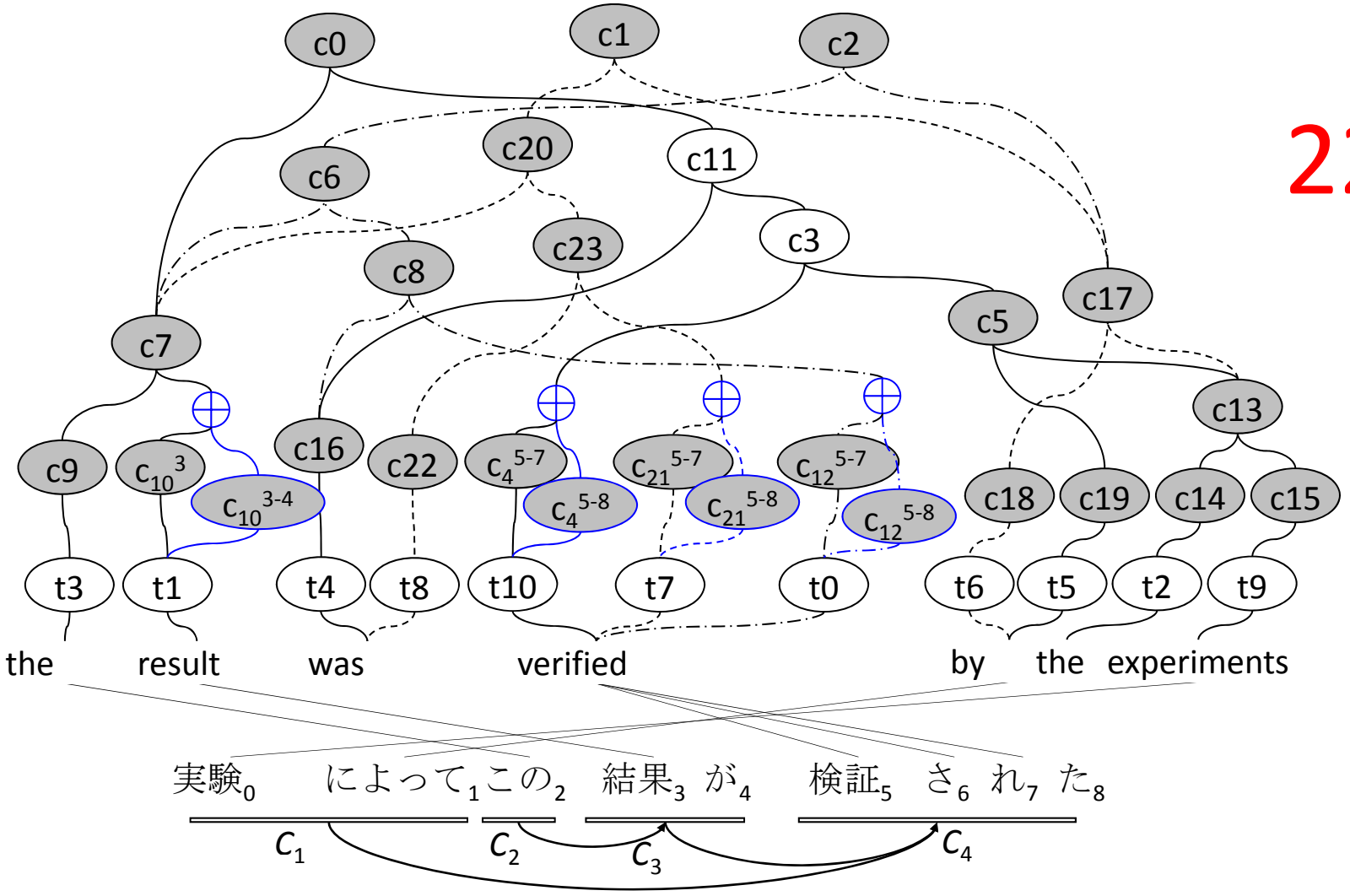


Additional comparison

Systems	En2Jp	Jp2En
Joshua (Li+, 2009)	21.79	19.73
Forest-based	**24.75****	**22.67****

**=p<0.01

Remove the ambiguous alignments



Remove the ambiguous alignments

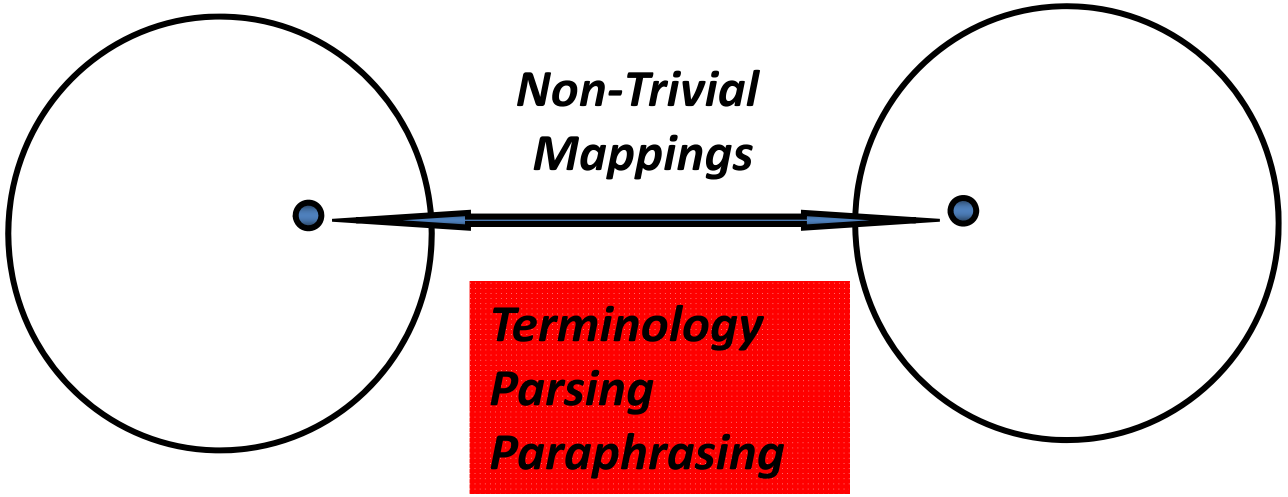
Systems	En2Jp
Forest-based rules (Wu+, ACL 2010)	24.75
+ forest-based decoder	25.83
+ function words	**28.44****

**=p<0.01

Plan of the Talk

- Backgrounds
  - Shifts in Society
  - Semantics-based information access to text and MT
- NLP and Semantics-based information access
  - Deep parsing and entity/event-based access
  - Efficient deep parsers
- MT with Deep Parsing
  - Structural Differences: Predicate-Argument Structures
  - Lexical Differences: Terms and Ontologies
- Concluding remarks and future work

From surface diversities and ambiguities to conceptual invariants



Language Domain

Linguistic expressions

Knowledge Domain

Concepts and Relationships among Them

Motivated Independently of language

Annotation of GENIA corpus – Term&POS

Event Annotation - Example