

Generic NLP technologies and domain adaptation for the analysis of biomedical text

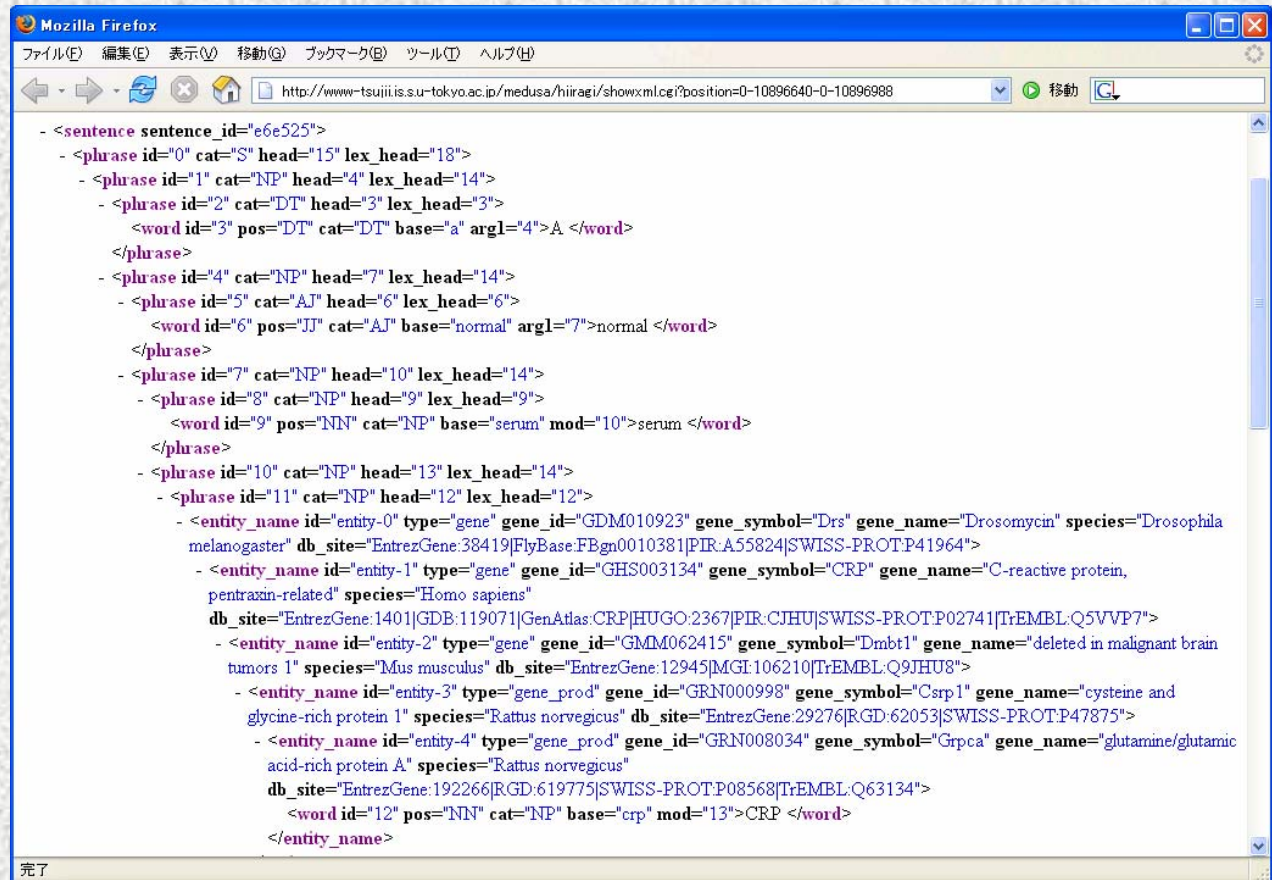
Yusuke Miyao
The University of Tokyo

Motivation

- Demands for efficient access methods for exploding text information
 - Research papers (e.g. MEDLINE)
 - Patent documents
 - Web
- Necessity of NLP technologies for the intelligent management of huge texts

Semantically annotated text

- “A normal serum **CRP** measurement does not exclude **deep vein thrombosis**”
- Annotated with syntactic/semantic structure and technical terms



```

- <sentence sentence_id="e6e525">
- <phrase id="0" cat="S" head="15" lex_head="18">
- <phrase id="1" cat="NP" head="4" lex_head="14">
- <phrase id="2" cat="DT" head="3" lex_head="3">
  <word id="3" pos="DT" cat="DT" base="a" arg1="4">A </word>
</phrase>
- <phrase id="4" cat="NP" head="7" lex_head="14">
- <phrase id="5" cat="AJ" head="6" lex_head="6">
  <word id="6" pos="JJ" cat="AJ" base="normal" arg1="7">normal </word>
</phrase>
- <phrase id="7" cat="NP" head="10" lex_head="14">
- <phrase id="8" cat="NP" head="9" lex_head="9">
  <word id="9" pos="NN" cat="NP" base="serum" mod="10">serum </word>
</phrase>
- <phrase id="10" cat="NP" head="13" lex_head="14">
- <phrase id="11" cat="NP" head="12" lex_head="12">
- <entity_name id="entity-0" type="gene" gene_id="GDM010923" gene_symbol="Drs" gene_name="Drosomycin" species="Drosophila melanogaster" db_site="EntrezGene:38419|FlyBase:FBgn0010381|PIR:A55824|SWISS-PROT:P41964">
- <entity_name id="entity-1" type="gene" gene_id="GHS003134" gene_symbol="CRP" gene_name="C-reactive protein, pentraxin-related" species="Homo sapiens" db_site="EntrezGene:1401|GDB:119071|GenAtlas:CRP|HUGO:2367|PIR:CJHU|SWISS-PROT:P02741|TrEMBL:Q5VVP7">
- <entity_name id="entity-2" type="gene" gene_id="GMM062415" gene_symbol="Dmbt1" gene_name="deleted in malignant brain tumors 1" species="Mus musculus" db_site="EntrezGene:12945|MGI:106210|TrEMBL:Q9JHU8">
- <entity_name id="entity-3" type="gene_prod" gene_id="GRN000998" gene_symbol="Csrp1" gene_name="cysteine and glycine-rich protein 1" species="Rattus norvegicus" db_site="EntrezGene:29276|RGD:62053|SWISS-PROT:P47875">
- <entity_name id="entity-4" type="gene_prod" gene_id="GRN008034" gene_symbol="Grpca" gene_name="glutamine/glutamic acid-rich protein A" species="Rattus norvegicus" db_site="EntrezGene:192266|RGD:619775|SWISS-PROT:P08568|TrEMBL:Q63134">
  <word id="12" pos="NN" cat="NP" base="crp" mod="13">CRP </word>
</entity_name>

```

Application: text mining

- Browsing protein interaction networks extracted from MEDLINE

The screenshot shows the CytoSailing web application in a Mozilla Firefox browser. The interface is divided into two main sections: the Interaction Viewer on the left and the Content Viewer on the right.

Interaction Viewer: This section allows users to drag interaction objects into evidence sentences. It features a table with columns for Entity, Object, and Match. The table lists various protein interactions, with Raf-1 and MAPK1 highlighted as the current focus.

Entity	Object	Match
RAF1	MAPK1	38 / 341
next	PKC	26 / 116
Erase	MAP2K1	23 / 147
	EGF	13 / 52
	AGT	12 / 39
	KSR	10 / 32
	PBP	10 / 30
	EPHB2	9 / 90
	OSM	9 / 15

Content Viewer: This section displays a list of sentences extracted from MEDLINE. The first sentence is selected, showing its full text and associated protein interactions. The sentence describes the activation of Raf-1 and its subsequent phosphorylation and activation of MEK1, which in turn phosphorylates and activates ERK1 and ERK2.

Sentences 1 -- 30 [Next](#)

1. PMID8557975 CATALYSIS
Active Raf-1 phosphorylates and activates the mitogen-activated protein (MAP) kinase/extracellular signal-regulated kinase kinase 1 (MEK1), which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, ERK1 and ERK2.
2. PMID8969227
Inhibition of protein kinase C (PKC) with calphostin C or down-regulation of PKC by pretreatment with a phorbol ester for 24 h abolished AngII-induced activation of Raf-1 and ERKs, and addition of a phorbol ester conversely induced a marked increase in the activities of Raf-1 and ERKs.
3. PMID10531364

Proteins that interact with "RAF1"

Application: text retrieval

- Retrieving sentences by specifying semantics

*Sentences
in which
“TNF induces
something”*

NLP tools for semantic annotation

- NLP tools are necessary for intelligent text mining/retrieval
- Requirements for NLP tools
 - The development of NLP tools must be inexpensive
 - NLP tools must compute accurate analyses of texts of the target domain

Our strategy

- Generic NLP tools are developed first
 - Large language resources are available for general domains (e.g. Penn Treebank)
- NLP tools are then adapted to specific domains
 - Exploring inexpensive methods for adaptation

Contents

- Introduction to our NLP tools
 - Part-of-speech tagger
 - Syntactic analyzer
 - Term recognizer
- Adaptation to the biomedical domain
 - Part-of-speech tagger
 - Syntactic analyzer

Part-of-speech tagger

- Annotating each word with its **part-of-speech tag**

The peri-kappa B site mediates human immunodeficiency
virus type 2 enhancer activation in monocytes ...

Det Noun Noun Noun Verb Adj Noun
Noun Noun Number Noun Noun Prep Noun

- First step for syntactic analysis and term recognition

Previous approach

- Sequence tagging models
 - Given word sequence $w_1 \dots w_n$, find the tag sequence $t_1 \dots t_n$ that maximizes the following probability

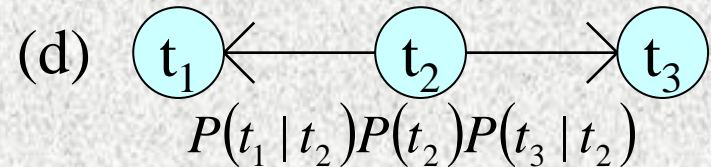
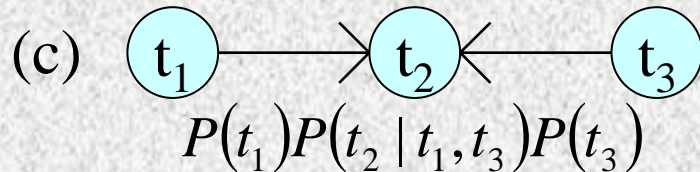
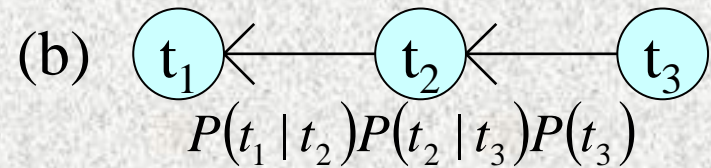
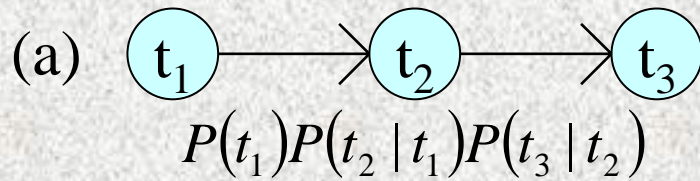
$$P(t_1 \dots t_n \mid w_1 \dots w_n)$$

- Left-to-right decomposition of the probability

$$P(t_1 \dots t_n \mid w_1 \dots w_n) \approx \prod_{i=1}^n P(t_i \mid t_{i-1} w_1 \dots w_n)$$

Our approach: bidirectional Inference

- Possible decomposition structures



- Bidirectional inference algorithm [Tsuruoka et al., 2005]

- We can find the “best” decomposition structure and tag sequences in polynomial time

Experimental results

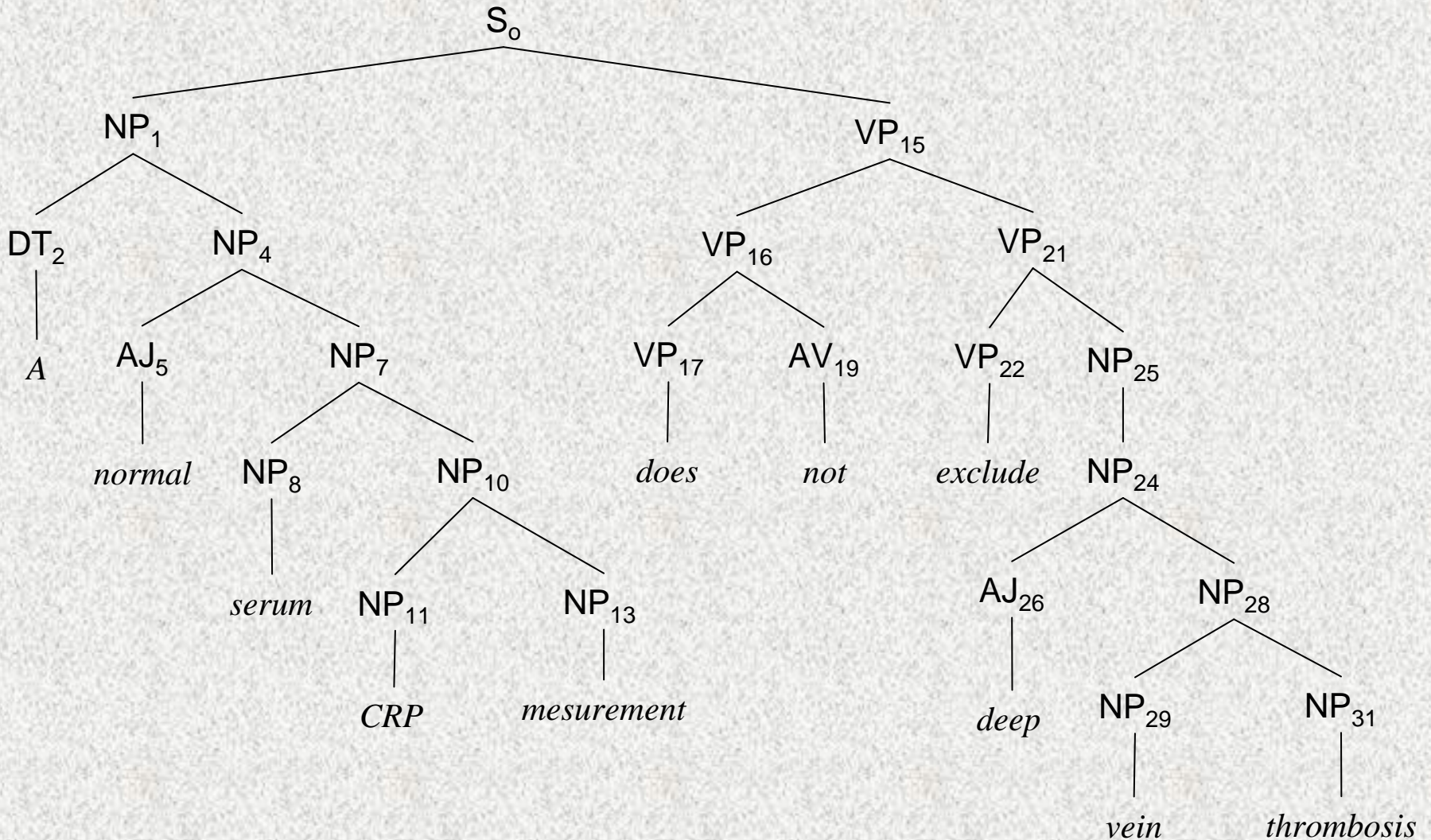
- Tagging speed and accuracy on Penn Treebank Wall Street Journal [Marcus et al., 1994]
 - Training set: Section 02-21 (39,832 sentences)
 - Test set: Section 23 (2,416 sentences)

	Tagging Speed	Accuracy
Dependency Net (2003)	Very slow	97.24
Perceptron (2002)	?	97.11
SVM (2003)	Fast	97.05
HMM (2000)	Extremely fast	96.48
Bidirectional inference	Very fast	97.10

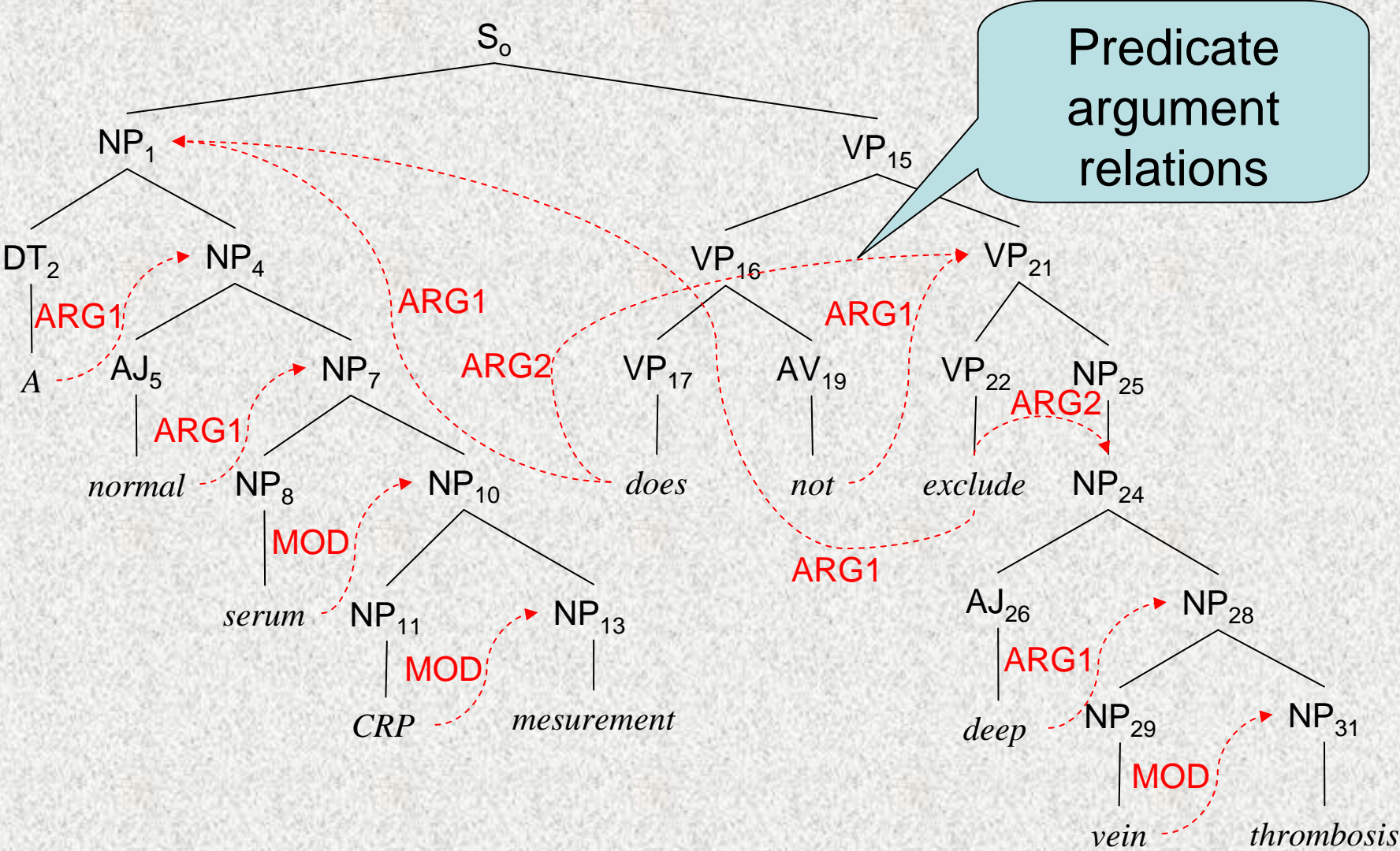
Syntactic analyzer

- Annotating sentences with their **parse trees** and **semantic structures**
- Necessary to obtain meanings expressed by various syntactic expressions

Parse tree



Semantic structure



Abstraction of surface expressions

Adenovirus-mediated high dose **p53** overexpression induced **Peg3 / Pw1** mRNA expression .

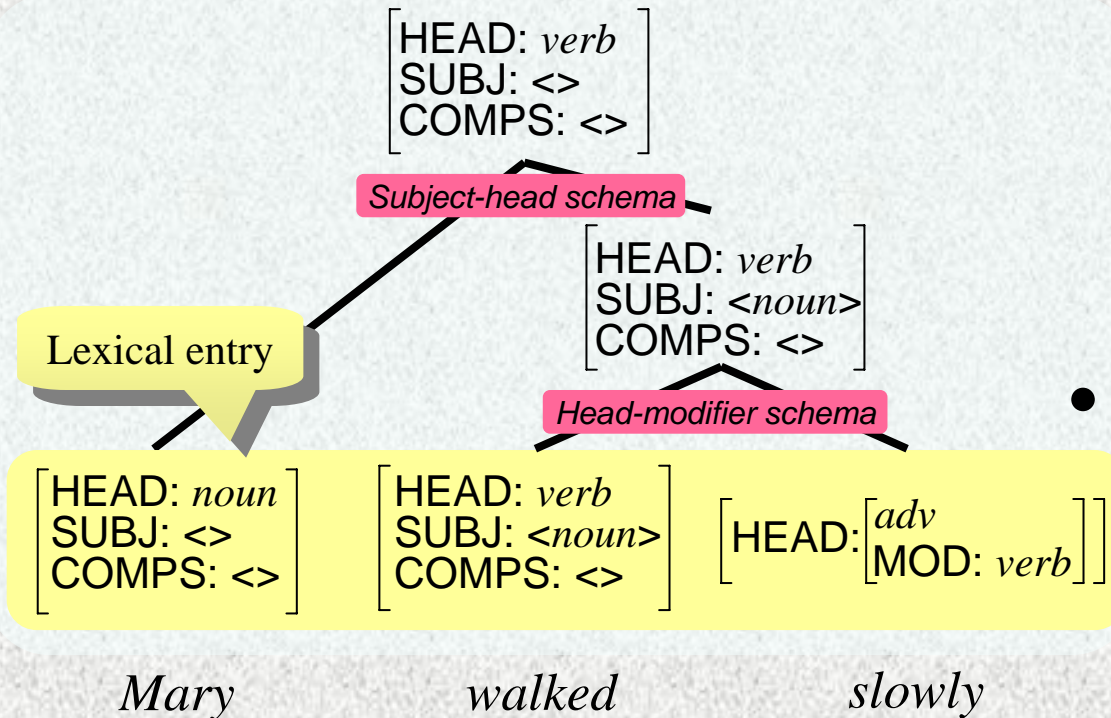
One of the mechanisms for **p53** to induce mitochondria-mediated **cell** death events is to activate genes that are directly involved in the initiation of mitochondria-induced apoptosis .

The **p53 gene** suppresses tumor **cell** growth by inducing **cell cycle arrest** or apoptosis .

Concomitant up-regulation of **p21** (**WAF1 / Cip1**) but not **p53** , especially in nodular hyperplasia , can be considered to induce **cell cycle arrest** of the parathyroid cells , but not cytotoxic effect of **OCT** .

Resistin overexpression is induced by a **beta3** adrenergic agonist in diet-related overweightness .

Our approach: HPSG parsing



- HPSG [Pollard and Sag, 1994] is a linguistic theory to explain sentence-to- semantics mappings
- Syntactic/semantic structures are computed with *grammar rules* and *lexical entries*

Development of an HPSG parser

- Wide-coverage HPSG grammar
 - Lexical entries are obtained from Penn Treebank [Miyao et al., 2004]
 - Probabilistic models of syntactic preferences are learned from Penn Treebank [Miyao and Tsujii, 2005]
- Parsing techniques
 - Local beam thresholding [Tsuruoka et al., 2004]
 - Global thresholding, iterative thresholding, etc. [Ninomiya et al., 2005]

Experimental results

- Training set: Penn Treebank Section 02-21 (39,832 sentences)
- Test set: Penn Treebank Section 23 (< 40 words, 2,164 sentences)
- Accuracy of predicate argument relations (i.e., *red arrows*) is measured

Precision	Recall	F-score	Avg. time (ms)
87.9%	86.9%	87.4%	360

Term recognizer

- Annotating phrases with **identifiers of ontological entities**
- e.x.) A normal serum CRP measurement does not exclude deep vein thrombosis

Class: *disease*
ID: *C0340708*
Name: *deep vein thrombosis*

Class: *protein*
ID: *GHS003134*
Name: *C-reactive protein*

- Mappings from textual expressions into ontology entries

Ontologies

- Ontologies have been developed in the biomedical domain
 - Gene Ontology, KEGG, UMLS, ICD, etc.
- E.g.) An entry from GENA [Koike & Takagi, 2004]

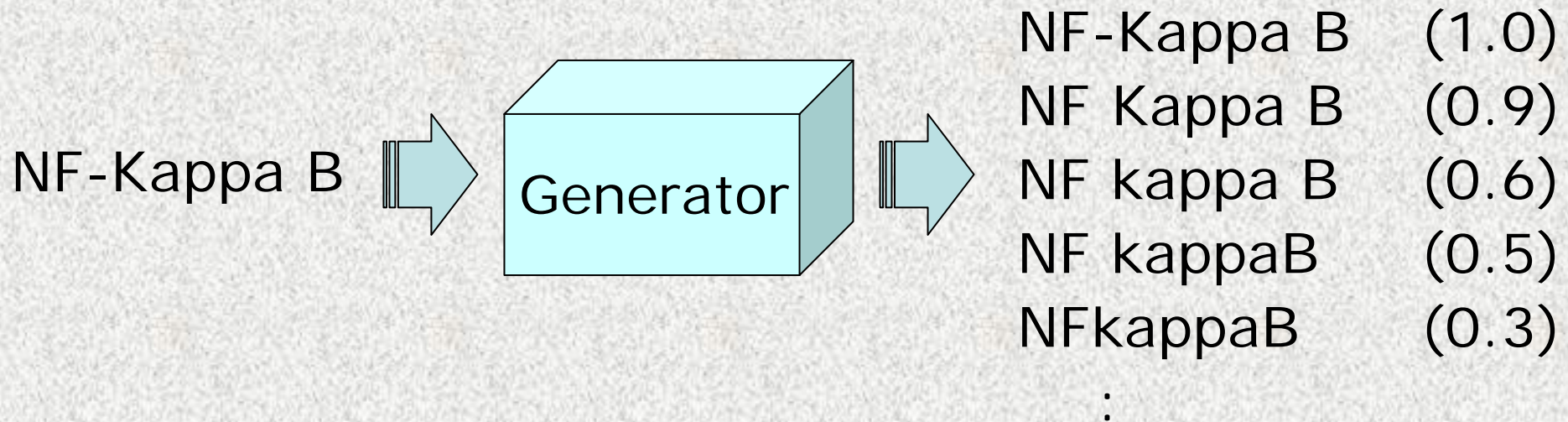
Symbol	CRP
Name	C-reactive protein, pentraxin-related
Species	Homo sapiens
Synonym	MGC88244, PTX1
Product	C-reactive protein precursor, C-reactive protein, pentraxin-related protein
External links	EntrezGene:1401, GDB: 119071, ...

Why term recognition difficult?

- Spelling variations complicate the problem
 - “Nuclear factor kappa B” is written as “NF kappa B”, “NF-kappa B”, “NFkappaB”, etc.
- Different surface expressions may denote the same entity in the ontology

Our approach: Automatic generation of spelling variants

- Variant Generator [Tsuruoka and Tsujii, 2003]



Each generated variant is associated with its generation probability

Example of variant generation (1)

Generation Probability	Generated Variants	Frequency
1.0 (input)	antiinflammatory effect	7
0.462	anti-inflammatory effect	33
0.393	antiinflammatory effects	6
0.356	Antiinflammatory effect	0
0.286	antiinflammatory-effect	0
0.181	anti-inflammatory effects	23
:	:	:

Example of variant generation (2)

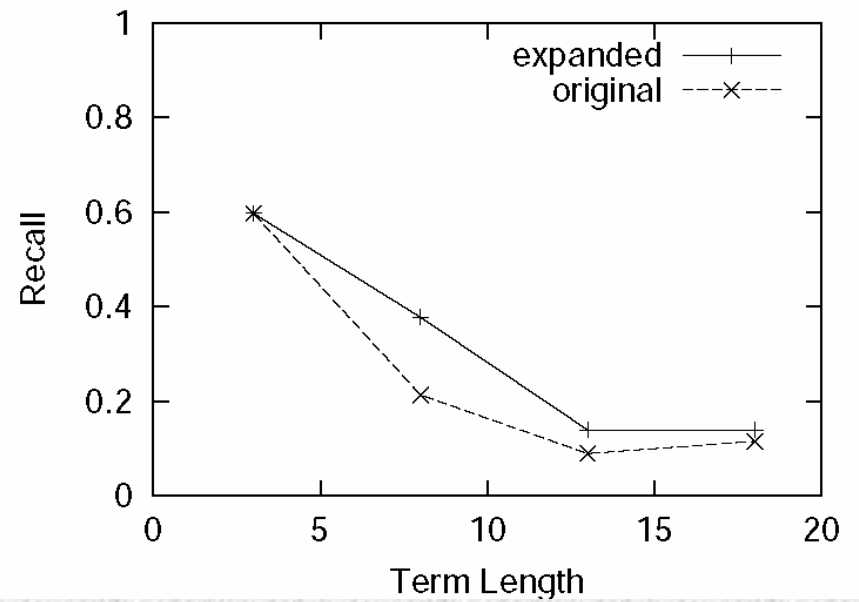
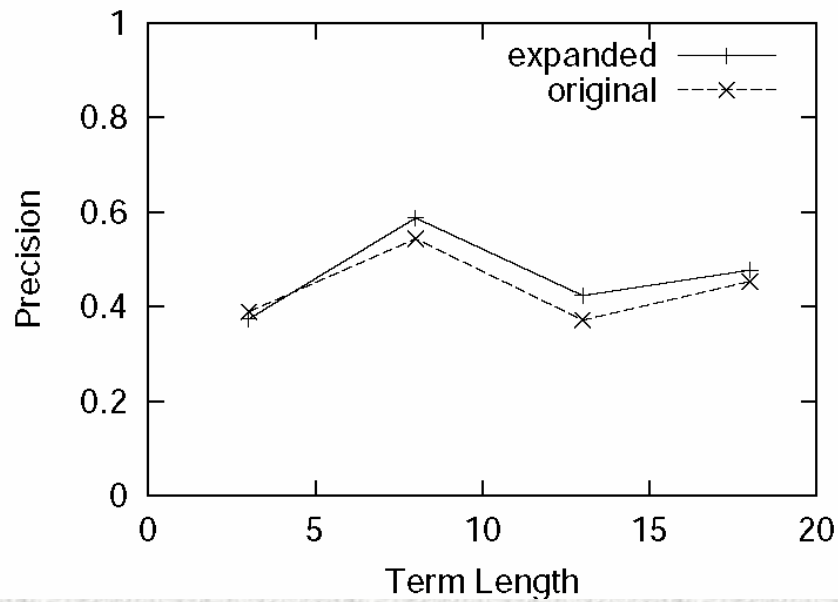
Generation Probability	Generated Variants	Frequency
1.0 (Input)	tumour necrosis factor alpha	15
0.492	tumor necrosis factor alpha	126
0.356	tumour necrosis factor-alpha	30
0.235	Tumour necrosis factor alpha	2
0.175	tumor necrosis factor alpha	182
0.115	Tumor necrosis factor alpha	8
:	:	:

Experiment:

Gene/Protein Name Recognition

- Corpus: GENIA NE corpus 3.01
 - For training: 1800 abstracts
 - For testing: 200 abstracts
 - Entity class: DNA, RNA, Protein
- Dictionary
 - Constructed from the GenBank database

Experimental results



Domain adaptation

- Large training data has been available for general domains (e.g. Penn Treebank WSJ)
- NLP Tools trained with general domain data are less accurate on the biomedical domain
- Development of domain-specific data requires considerable human efforts

Tagging errors by TnT tagger [Brants, 2000]

... 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 ~~RR~~ ~~JJ~~ NNS IN

Stimulation of interferon beta gene transcription in vitro by

NN IN JJ JJ NN NN ~~IN~~ ~~NN~~ IN

- Accuracy of the TnT tagger on the GENIA POS corpus: 84.4%

Our approach: re-training of maximum entropy models

- Our part-of-speech tagger and syntactic analyzer are trained as *maximum entropy models* [Berger et al., 2000]
 - Model parameters are automatically estimated so as to maximize the likelihood of training data

Model parameter **Feature function**
(given by the developer)

$$p(x) = \frac{1}{Z} \exp\left(\sum_{i=1}^F \lambda_i f_i(x)\right)$$

- Adapting maximum entropy models to target domains by re-training with domain specific data

Methods for domain adaptation

- **Combined training data:** a model is trained from scratch with the original and domain-specific data
- **Reference distribution:** an original model is used as a *reference probabilistic distribution* of a new model

$$p_{new}(x) = \frac{1}{Z} p_{orig}(x) \exp\left(\sum_{i=1}^F \lambda_i f_i(x)\right)$$

Adaptation of the part-of-speech tagger

- Relationships between training and test data are evaluated with the following corpora
 - **WSJ**: Penn Treebank WSJ
 - **GENIA**: GENIA POS corpus [Kim et al., 2003]
 - 2,000 MEDLINE abstracts selected by MeSH terms, *Human*, *Blood cells*, and *Transcription factors*
 - **PennBioIE**: Penn BioIE corpus [Kulick et al., 2004]
 - 1,100 MEDLINE abstracts about inhibition of the cytochrome P450 family of enzymes
 - 1,157 MEDLINE abstracts about molecular genetics of cancer
 - **Fly**: 200 MEDLINE abstracts on *Drosophila melanogaster*

Training and test sets

- Training sets

	# tokens	# sentences
WSJ	912,344	38,219
GENIA	450,492	18,508
PennBioIE	641,838	29,422
Fly	24,450	1,024

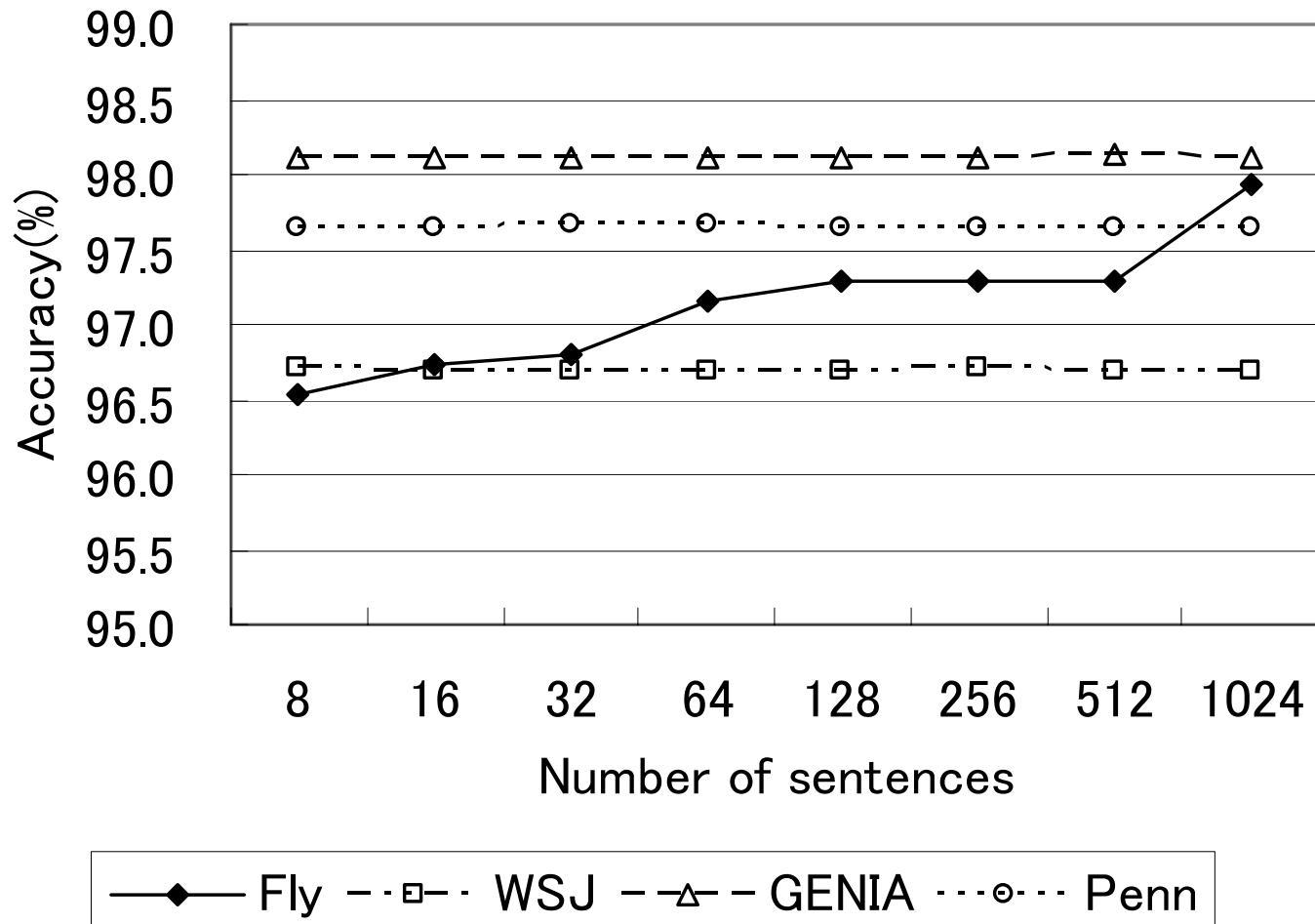
- Test sets

	# tokens	# sentences
WSJ	129,654	5,462
GENIA	50,562	2,036
PennBioIE	70,713	3,270
Fly	7,615	326

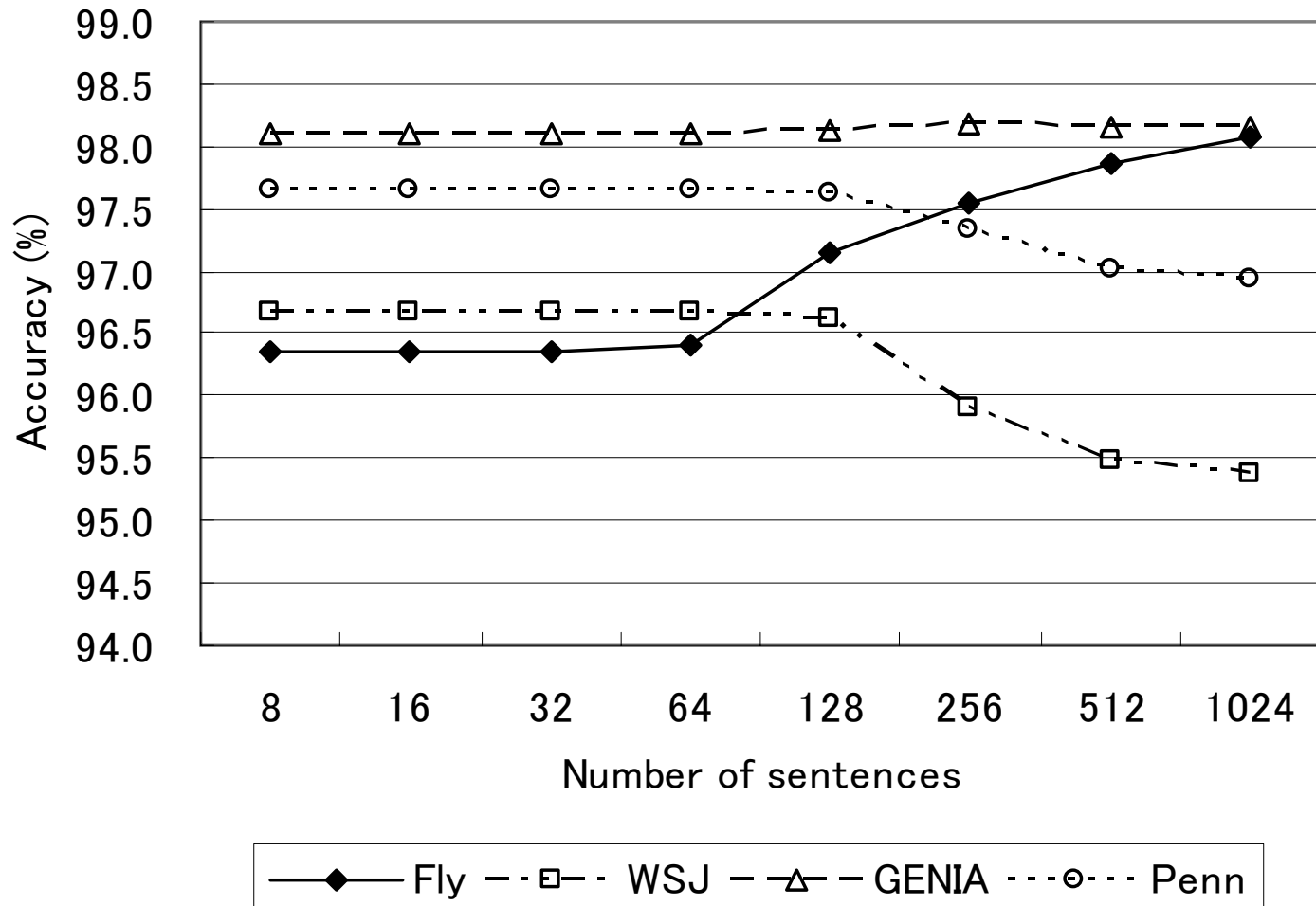
Experimental results

	Accuracy				Training time (sec.)
	WSJ	GENIA	PennBioIE	Fly	
WSJ+GENIA+PennBioIE	96.68	98.10	97.65	96.35	
Fly only				93.91	
Combined	96.69	98.12	97.65	97.94	30,632
Ref. dist	95.38	98.17	96.93	98.08	21

Corpus size vs. accuracy (combined training data)



Corpus size vs. accuracy (reference distribution)



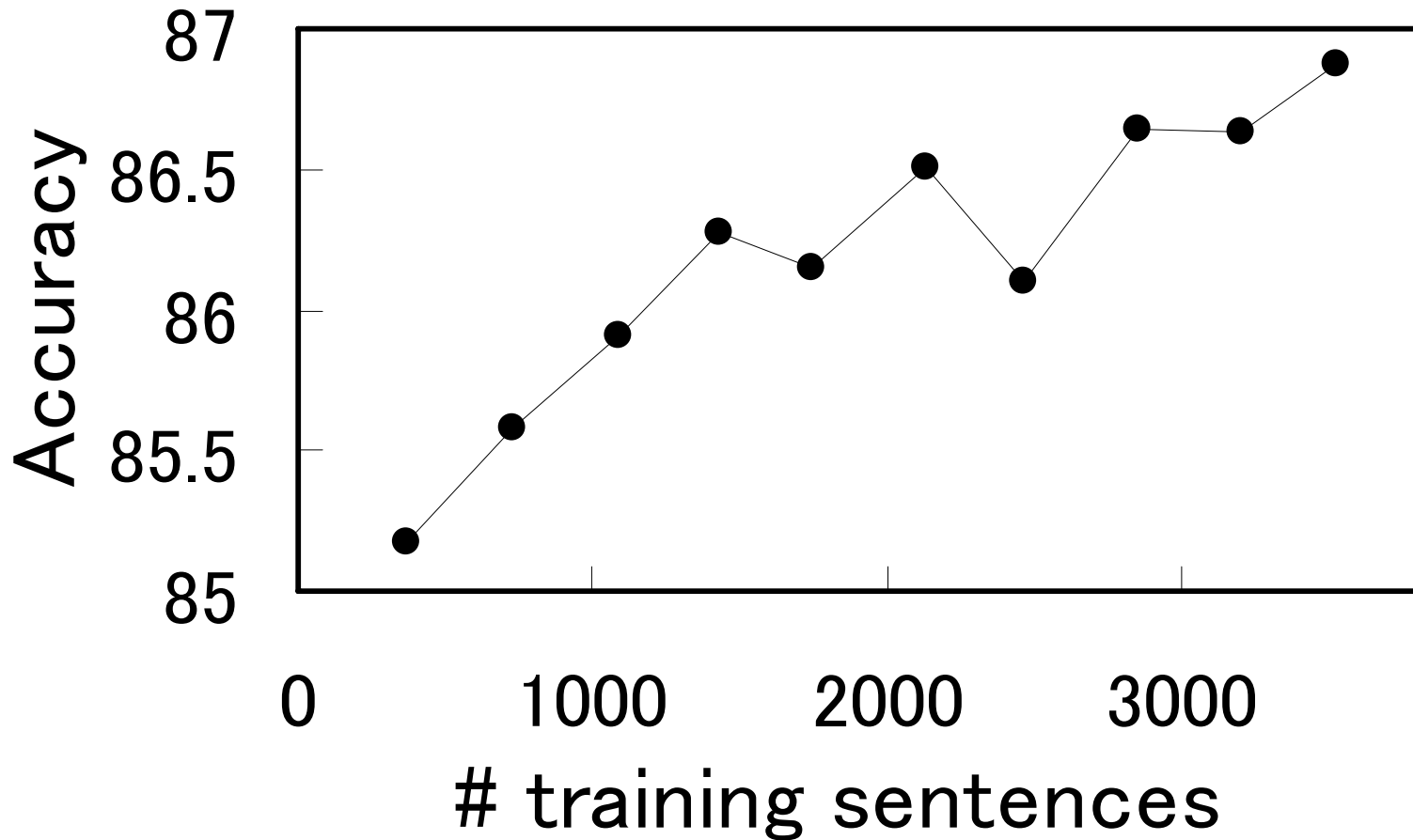
Adaptation of the syntactic analyzer

- The same methods were applied to our syntactic analyzer [Hara et al., 2005]
- Training set:
 - **WSJ**: Penn Treebank WSJ 39,832 sentences
 - **GENIA**: GENIA Treebank 3,524 sentences
- Test set:
 - **WSJ**: Penn Treebank WSJ 2,164 sentences
 - **GENIA**: GENIA Treebank 467 sentences

Experimental results

	Accuracy		Training time (sec.)
	WSJ	GENIA	
WSJ only	87.16	85.10	137,038
GENIA only	42.49	85.72	1,694
Combined	86.09	86.32	29,421
Ref. dist	86.81	86.87	2,278

Corpus size vs. Accuracy



Summary

- NLP tools for general domains have been developed using large training data
- These tools have successfully been adapted to the biomedical domain using small training data
- Accurate NLP tools are available for the biomedical domain
 - Part-of-speech tagger
 - Syntactic analyzer
 - Term recognizer

Products

- Our software products are available online
 - **uptagger**: a part-of-speech tagger for general domains
 - **geniatagger**: a part-of-speech tagger for the biomedical domain
 - **enju**: a syntactic analyzer based on HPSG
- See our homepage for details
 - <http://www-tsujii.is.s.u-tokyo.ac.jp/>