What's in a name?

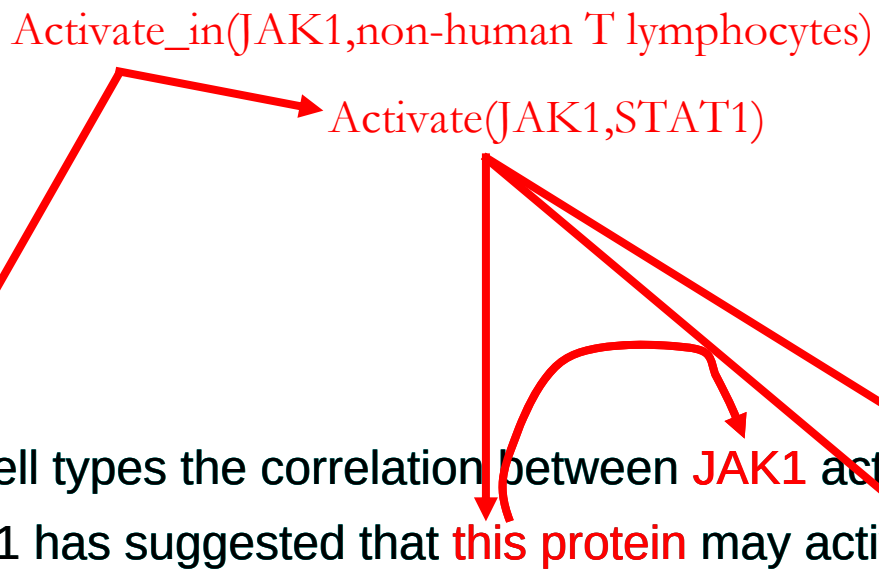
Nigel Collier
Associate Professor
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430

www.research.nii.ac.jp/~collier
collier@nii.ac.jp

Outline

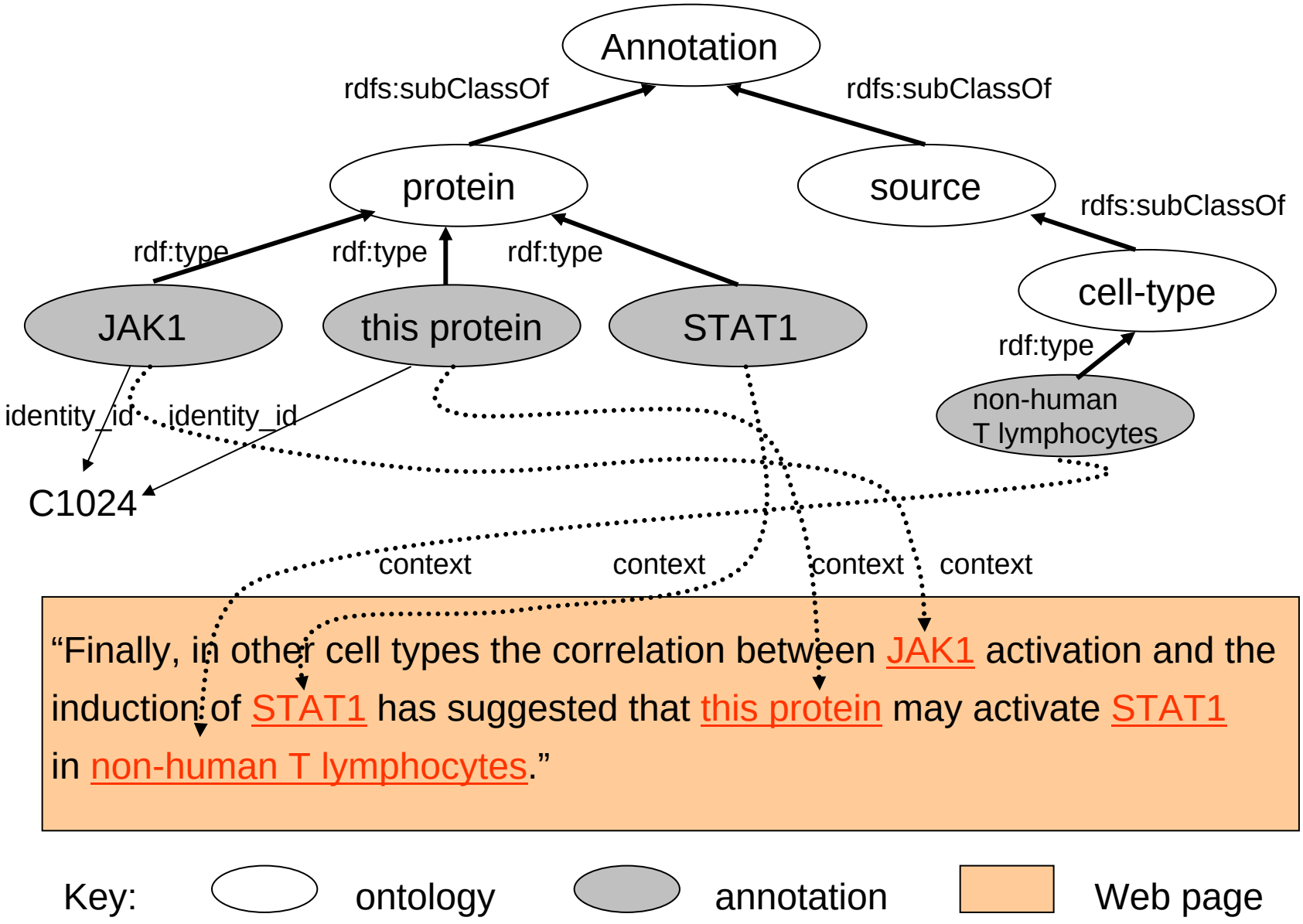
- Introduction
  - An overview of the text mining task
- The nature of biological entities
  - The state of the art
  - Challenges
- Training computers to do bio-NE
  - Knowledge acquisition
  - Methodology
  - Results from JNLPBA
- Discussion and future work

Introduction

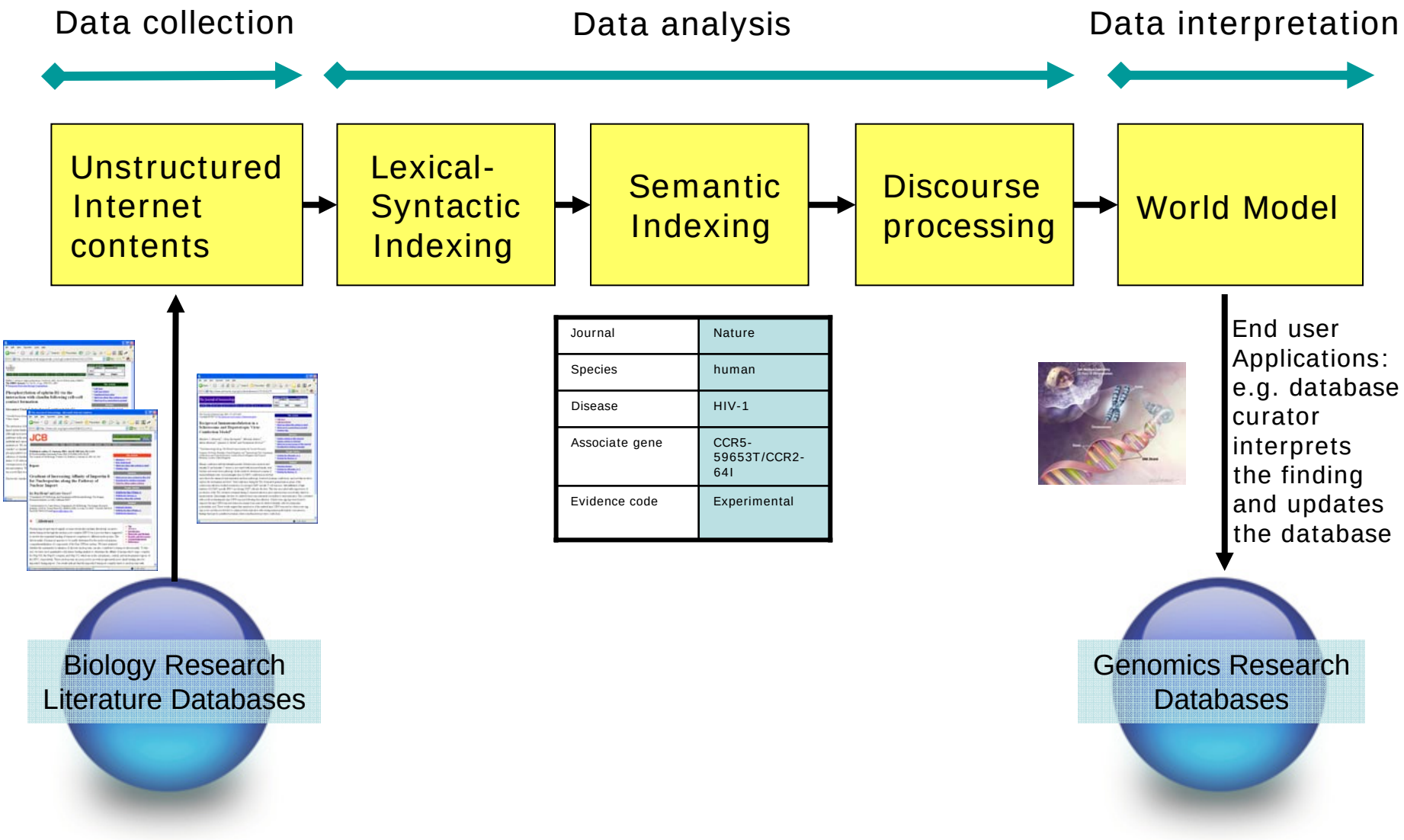


"Finally, in other cell types the correlation between JAK1 activation and the induction of STAT1 has suggested that this protein may activate STAT1 in non-human T lymphocytes."

Ontology		Instances		
Taxonomy	Axioms	Concepts	Coreference	Relations
Activate(JAK1,STAT1), Activate_in(JAK1,non-human T lymphocytes)				

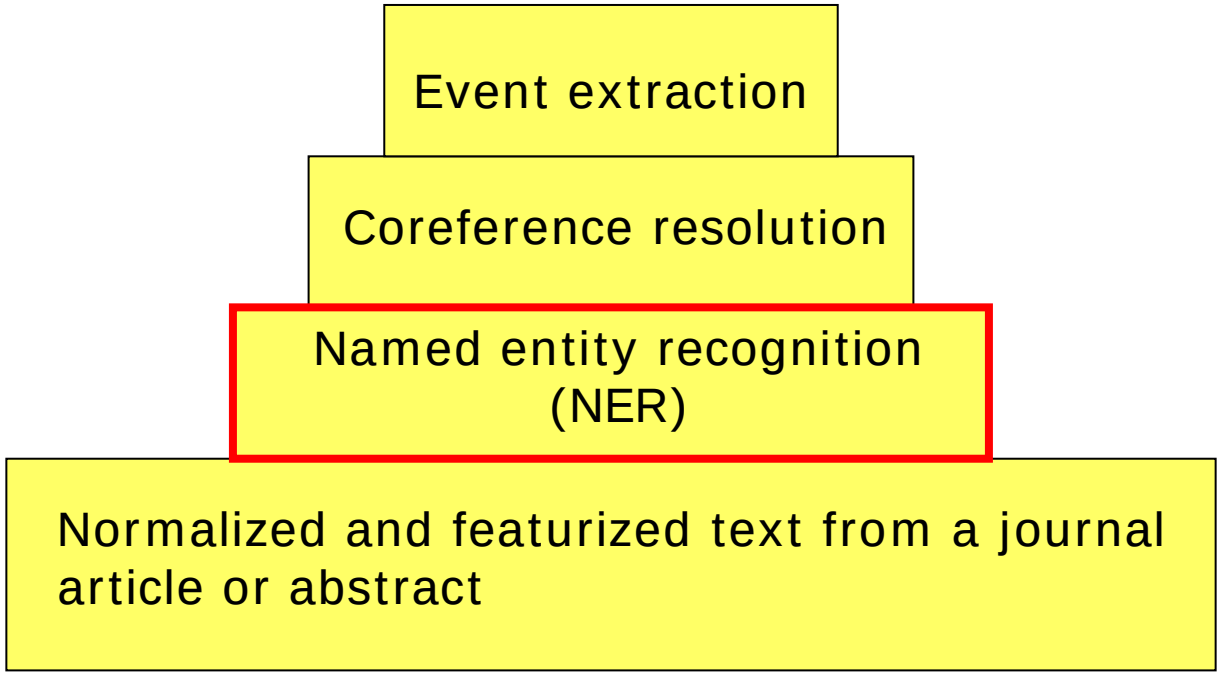

Annotation
rdfs:subClassOf
rdfs:subClassOf
protein
source
rdfs:subClassOf
rdf:type
rdf:type
rdf:type
cell-type
JAK1
this protein
STAT1
rdf:type
non-human
T lymphocytes
identity_id
identity_id
C1024
context
context
context
context
“Finally, in other cell types the correlation between JAK1 activation and the induction of STAT1 has suggested that this protein may activate STAT1 in non-human T lymphocytes.”
Key:
ontology
annotation
Web page

The Pipeline Architecture



Tasks in semantic indexing

NE task definition: given a document (full article or abstract), identify non-overlapping sequences of word tokens and assign them an entity class representing the concepts of interest.



Why is NER necessary?

- New terms are being invented all the time

"We have previously identified a J binding protein (JBP1) involved in propagating J synthesis. We have now identified a homolog of JBP1, JBP2, containing a domain related to the SWI2/SNF2 family of chromatin remodeling proteins that is upregulated in bloodstream form cells and interacts with nuclean chromatin"

[DiPaolo, C., Kieft, R., Cross, M. and Sabatini, R. Mol. Cell 17(3) 2005]

- Biological databases are not up to date and do not list all variant forms

The nature of biological entities

The JNLPBA 2004 shared task

- Systematic evaluation of entity tagging by machines against a human gold standard
- Domain was ‘human’ ‘blood cell’ ‘transcription factor’
- Open system task – entrants were free to use whatever resources they could think of in addition to the GENIA corpus (~2000 human annotated abstracts);
- Common evaluation of machine capability on a human annotated gold standard (~400 newly annotated MEDLINE abstracts)

Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y. and Collier, N. (2004), "Introduction to the Bio-Entity Recognition Task at JNLPBA", in proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 28-29 August, Geneva, Switzerland.

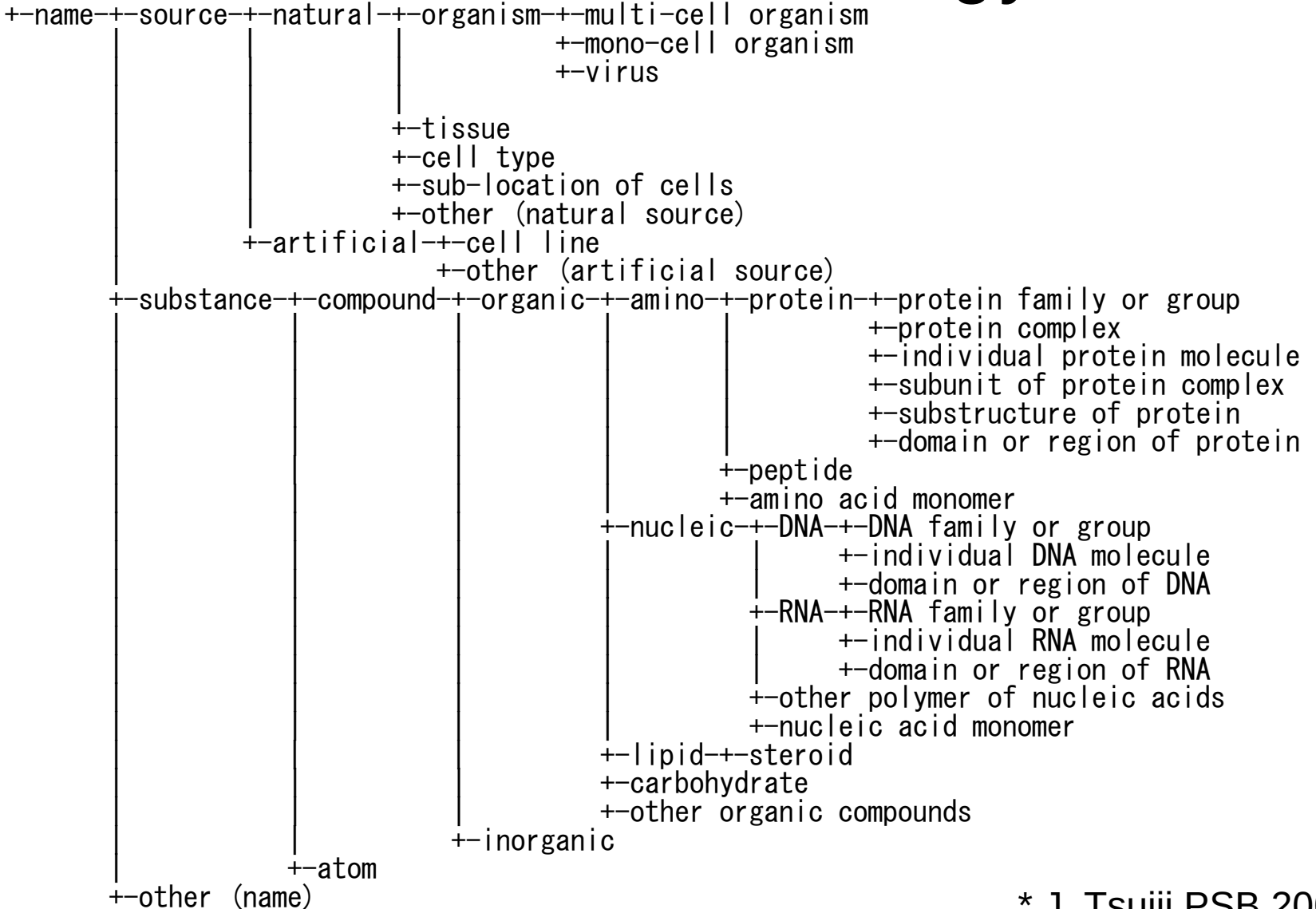
Example from the JNLPBA shared-task

A complete inhibition of DNA synthesis by dexamethasone (Dx) could be observed when IL 2-depleted cultures of CTL were either incubated for 6 h with the hormone prior to the addition of IL 2 or treated simultaneously with Dx and a low concentration of IL 2.

cell line
protein

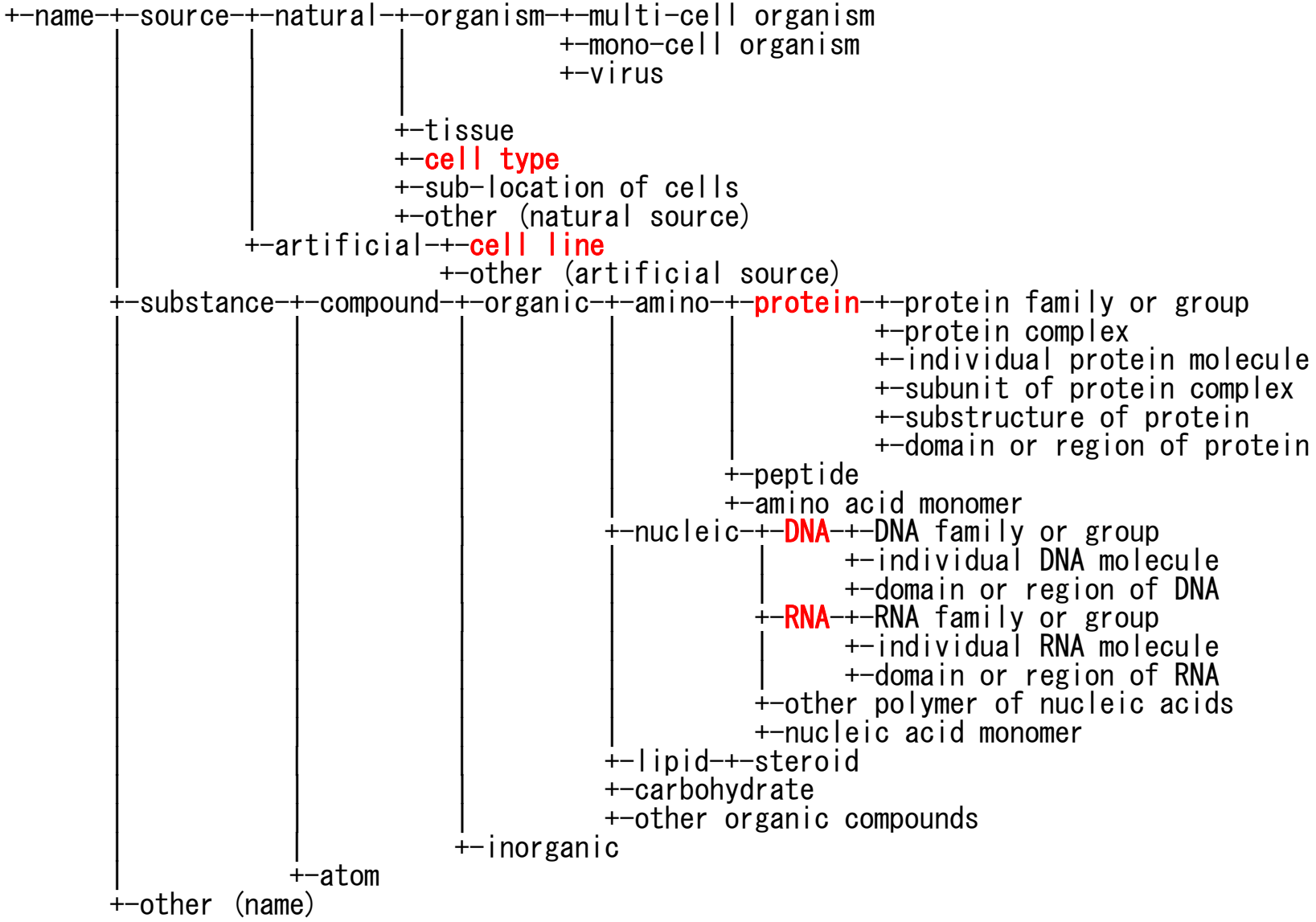
- The task takes a particular view of the ontology which is suitable for tagging non-overlapping spans of text.
- Nevertheless compromises need to be made at both the schema level and in the annotation guidelines.
  - e.g. “dexamethasone” is ignored
  - e.g. “IL 2” is not tagged as a protein inside “IL 2-depleted cultures”

The GENIA ontology



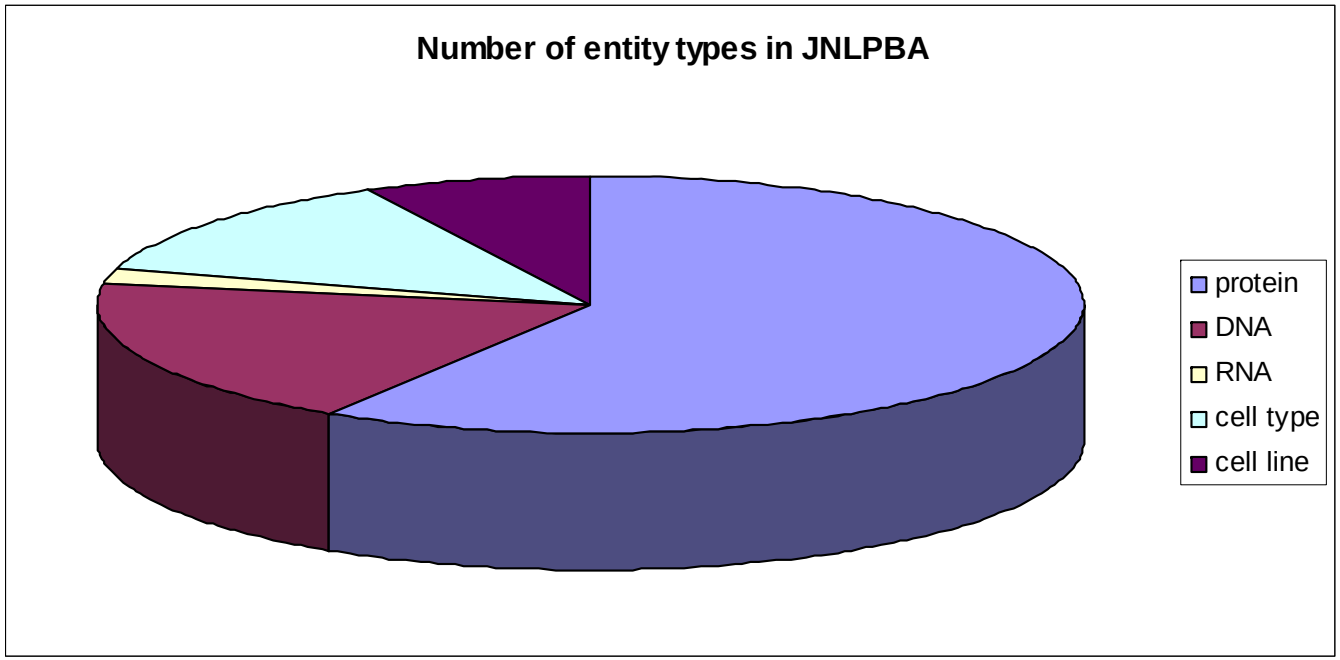
* J. Tsujii PSB 2001

JNLPBA shared task classes



Test collection characteristics

- 20,546 sentences, 472,006 tokens
- Named entity counts: 30,269 protein (15.1%), 9533 DNA (4.8%), 951 RNA (0.5%), 6718 cell type (3.4%), 3830 cell line (1.9%)



Evaluation Metrics

- Precision and Recall
  - Precision: Correct answer/Answers produced
  - Recall: Correct answers/Total possible correct answers
- F-measure
  - Where β is a parameter representing relative importance of P and R

$$F = \frac{(\beta^2 + 1)PR}{(\beta^2 P + R)}$$

Experience in shared-evaluations

- The state-of-the-art in news entity tagging is at ‘near human’ levels of performance – high 90s F-score (e.g. MUC 1995, CoNLL 2003);
- The state of the art for bio-entity tagging in JNLPBA shared evaluation task is in the mid-70s

Challenges in biological entity recognition [1]

- Term variant forms
  - Orthographic variants (e.g. T cell, t cell | Interleukin-2, Interleukin 2)
  - Use of capitalization and hyphenation is idiosyncratic
  - Morphological variants (e.g. protein, proteins | anti-CD28, CD28)
  - Aliases and abbreviations (e.g. human immunodeficiency type 2, HIV-2)
- Descriptive naming
  - e.g. normal thymic epithelial cells [Zhou et al. 2003]
- Uncontrolled naming
  - Experience in BioCreative 1b tagging of gene names in model organisms confirmed that fly genes were far more difficult than yeast or mouse.

Challenges in biological entity recognition [2]

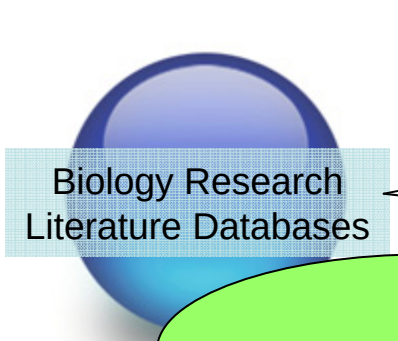
- Name length
  - e.g. 47 kDa sterol regulatory element binding factor 18.6% of NEs in GENIA v3.0 have >= length 4 [Zhou et al. 2003]
  - Average gene name is 2.09 in BioCreative 1a compared to 1.69 for organization names in MUC-6
- Syntactic variants
  - Conjunction and disjunction (e.g. c- and v-rel (proto) oncogenes)
  - 2.1% of NEs in GENIA v3.0
- Semantic ambiguity
  - Due to context (e.g. interleukin-2 as PROTEIN or DNA)
  - Due to granularity (e.g. interleukin-2 gene expression as OTHER_NAME)

Challenges in biological entity recognition [3]

- Widespread use of abbreviation
  - e.g. APC as activated protein c, aphidicholin, atrial premature complexes, adenomatous polyposis coli, antigen presenting cells [Tsuruoka et al. 2003]
  - Challenging cases, e.g. GNAT as Gcn5-related N-acetyltransferase [Schwartz & Hearst 2003]

Training computers to do bio-NE

Experimental framework



Knowledge Markup

Text Normalizer

- 16 node cluster computer with 64 Intel Xeon CPUs
- Snap Server 18000 5.4 TB disk array (expandable) storage system
- Sun Grid Engine
- All nodes connected by GB Ethernet
- Operational and expanding since 2002
- Integrated NLP tools, algorithms and resources in a common data model

Wizard cluster

Feature Composer

mode=train|install

mode=test

Train SVM

SVM Model

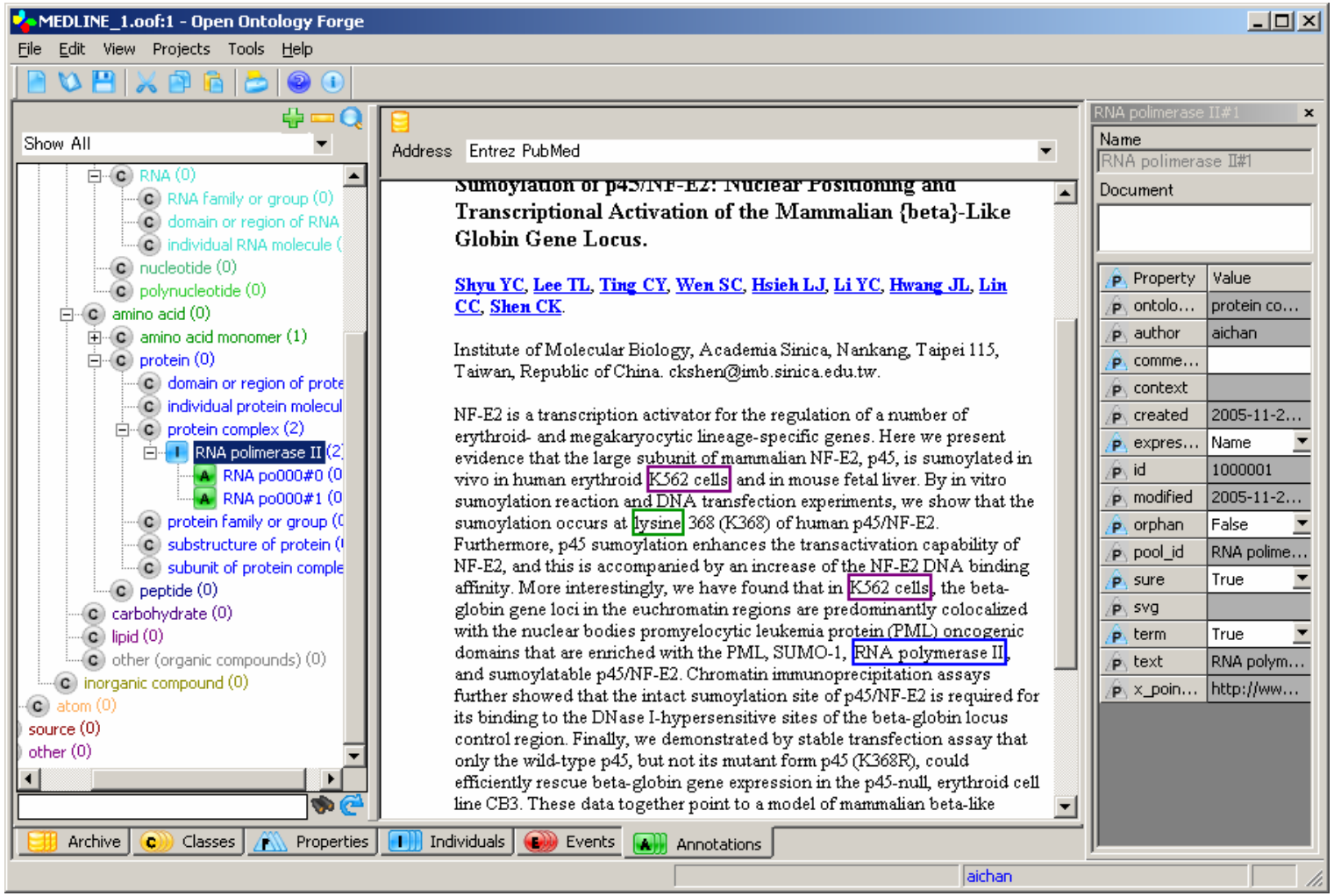
Test SVM

Tagged Document

Evaluate

F-scores

Knowledge markup with OOF



Kawazoe, A., Kitamoto, A. and Collier, N. (2004), in proceedings LREC'2004, Lisbon, Portugal.

Major features of OOF

- Handles large document collections using internal archiving of documents with MySQL database
- Simple process of ontology creation
  - Support for taxonomies, classes, properties, individuals and annotations
- Annotation of text/image with linkage to ontologies
- Annotation of pooled coreference relations using referred individuals
- Three formats for ontology/annotation export
  - RDF(S)
  - In-line XML
  - HTML
- Version 2 release from December (http://research.nii.ac.jp/~collier) – end of advertising!

Feature types

Knowledge type	Feature name
Surface text	Word features
Orthographic	Othographic features
Morphological	Prefix/suffix
	Part of speech (POS)
	Lemma
Syntactic	Parenthesis matching
	Head of noun phrase
	Predicate-argument relations
Semantic	Previous NE tags
Discourse	Abbreviation full forms

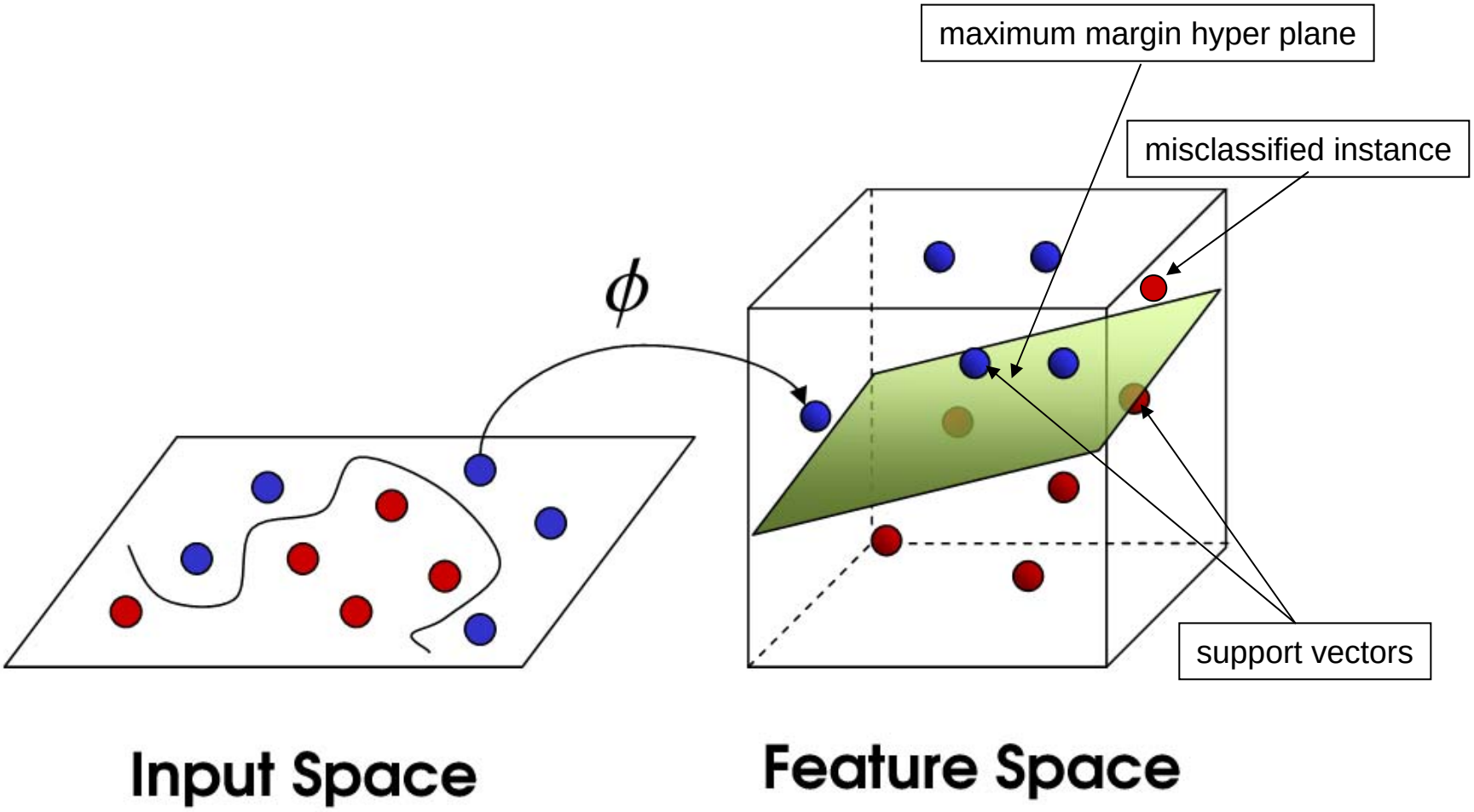
Examples of features

1. Activation activation Activation - - N NOM_SG A231 ic O
2. of of - activation mod PREP - O100 lw O
3. JAK jak kinases kinase attr N NOM_SG J200 03 B-PROTEIN
4. kinases kinase kinases of pcomp N NOM_PL K522 04 I-PROTEIN
5. and and - - - CC - A530 lw O
6. STAT stat proteins protein attr ABBR NOM_SG S330 03 B-PROTEIN
7. proteins protein proteins - - N NOM_PL P635 04 I-PROTEIN
8. by by - protein mod PREP - B000 lw O
9. interleukin-2 interleukin-2 alpha interferon attr N NOM_SG I536 11 B-PROTEIN
10. and and alpha interferon attr CC - A530 lw O
11. interferon interferon alpha alpha attr N NOM_SG I536 lw B-PROTEIN
12. alpha alpha alpha by pcomp N NOM_SG A410 01 I-PROTEIN
13. , , - - - - - - cm O
14. but but - - - CC - B300 lw O
15. not not - - - NEG-PART - N300 lw O
16. the the - - - DET - T000 lw O
17. T t receptor cell attr ABBR NOM_SG T000 06 B-PROTEIN
18. cell cell receptor antigen attr N NOM_SG C400 05 I-PROTEIN
19. antigen antigen receptor receptor attr N NOM_SG A532 lw I-PROTEIN
20. receptor receptor receptor - - N NOM_SG R213 04 I-PROTEIN
21. , , - - - - - - cm O

SVM Model

- Based on work of Vapnik 1995
- Most popular model used in JNLPBA 5 out of 8 systems (others were HMM, MEMM, CRF)
- Non-probabilistic classifier
  - Maximum margin hyperplane
  - Use of kernel functions to perform non-linear classification with minimal computational cost
- Robust to noise – can ignore outliers
- Multi-classifiers are built up from binary classifiers
- Achieved state-of-the-art performance in many classification tasks

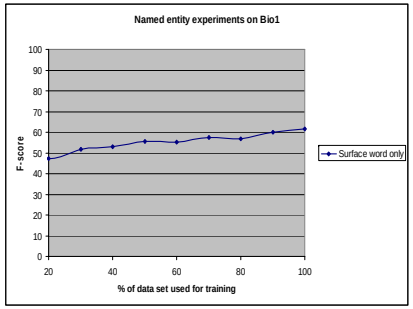
Classification in SVMs

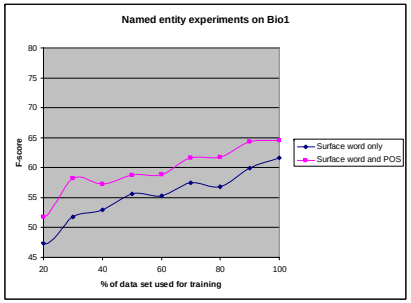


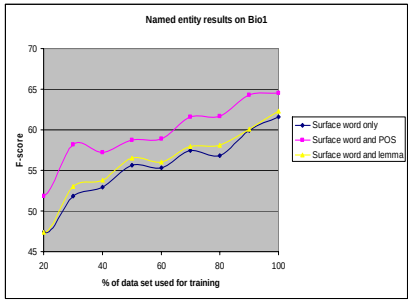
SVM Features

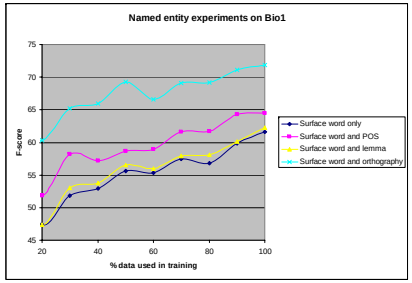
Word	Activation	of	JAK	kinases	and	STAT	proteins	by
POS	N	PREP	N	N	CC	ABBR	N	PREP
Orthography	ic	lw	uc	lw	lw	uc	lw	lw
Class	O	O	B-PROTEIN	I-PROTEIN	O	B-PROTEIN	B-PROTEIN	O

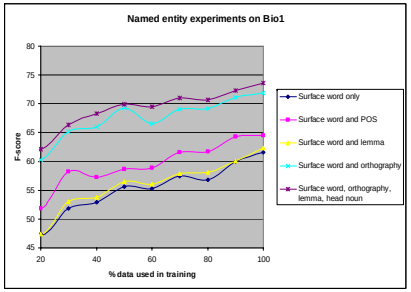
Feature farming

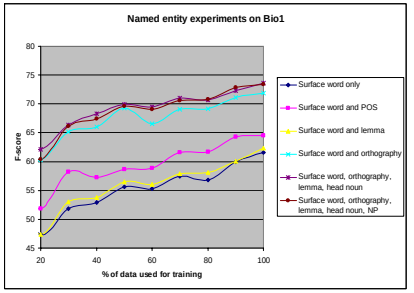


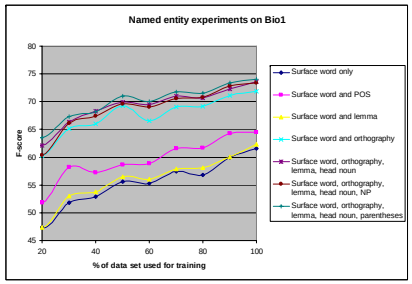


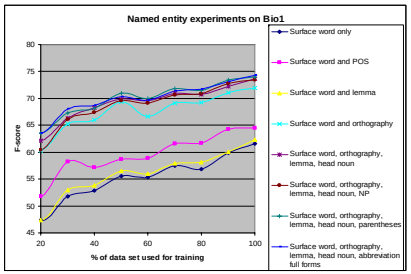


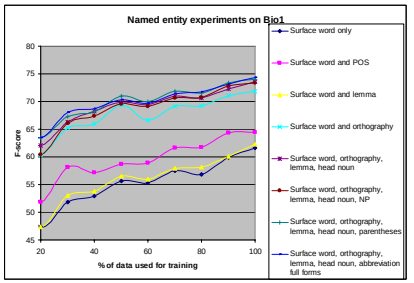






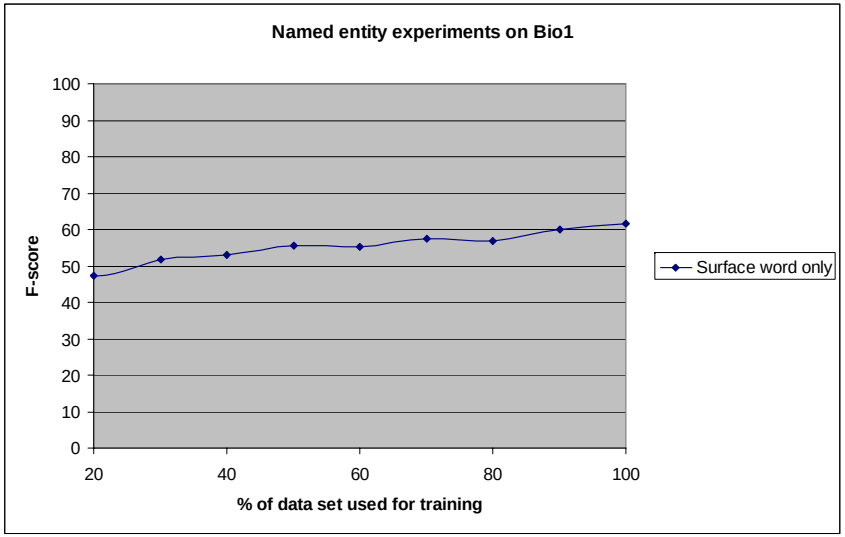






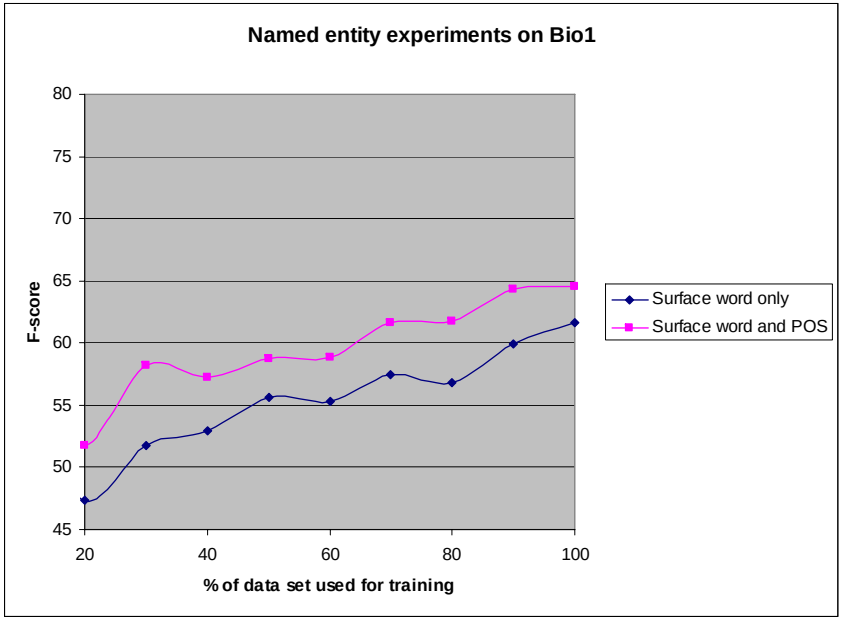
2 data sets x 10 way split of the data set x 10 fold cross validation x 9 feature splits = 1800 experiments in about 24 hours

Results [1]: Surface words



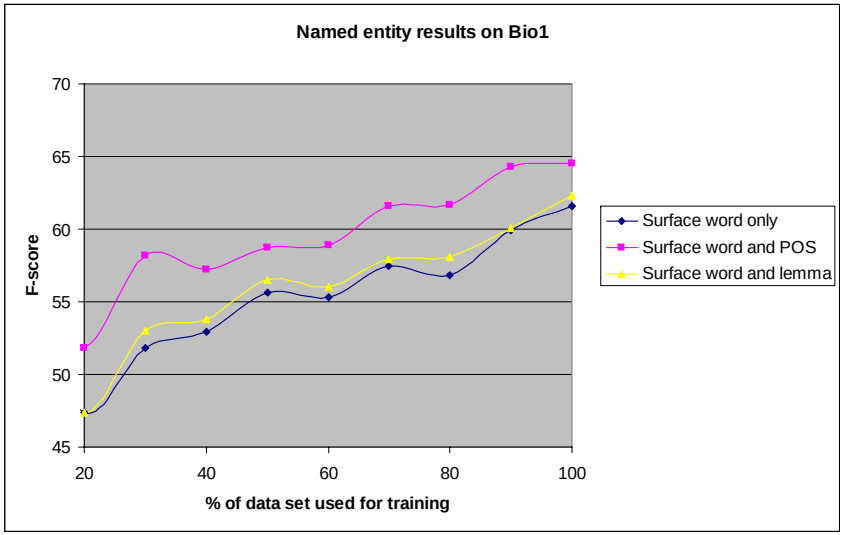
- Intuition
  - We've seen:
  - $[\text{JAK kinase}]_{\text{protein}}$
  - and
  - $[\text{Jun}]_{\text{protein}}$
  - Guess that:
  - $[\text{Jun kinase}]_{\text{protein}}$

Results [2]: Words plus POS



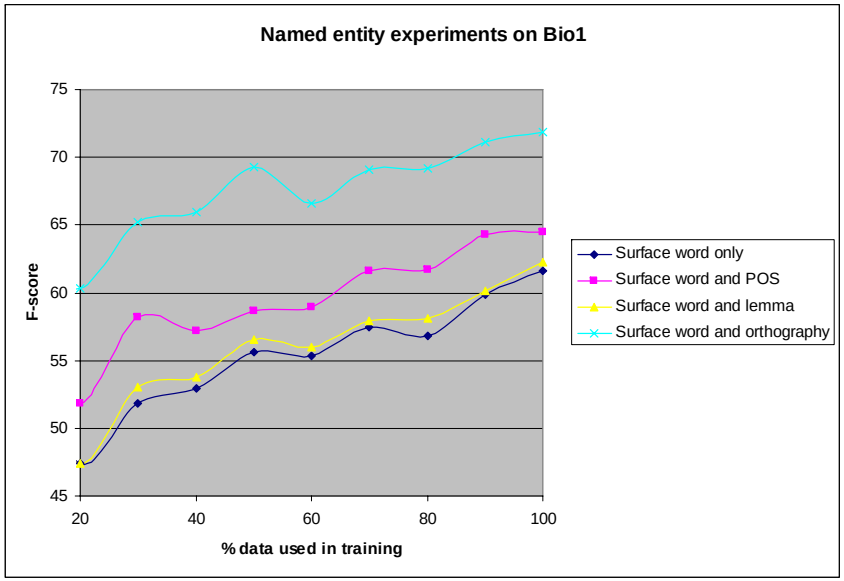
- Intuition
  - We've seen:
  - $[\text{JAK\_N kinase\_N}]_{\text{protein}}$
  - Hypothesize that:
  - $[\text{?\_N kinase\_N}]_{\text{protein}}$

Results [3]: Words and lemma



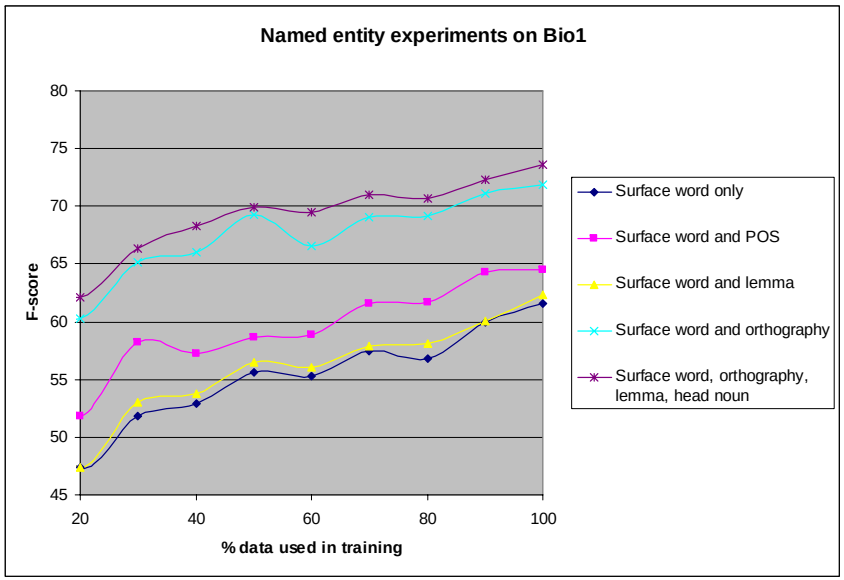
- Intuition
  - We've seen:
  - [JAK_JAK kinases_kinase]$_{protein}$
  - Hypothesize that:
  - [JAK_JAK kinase_kinase]$_{protein}$

Results [4]: Words and orthography



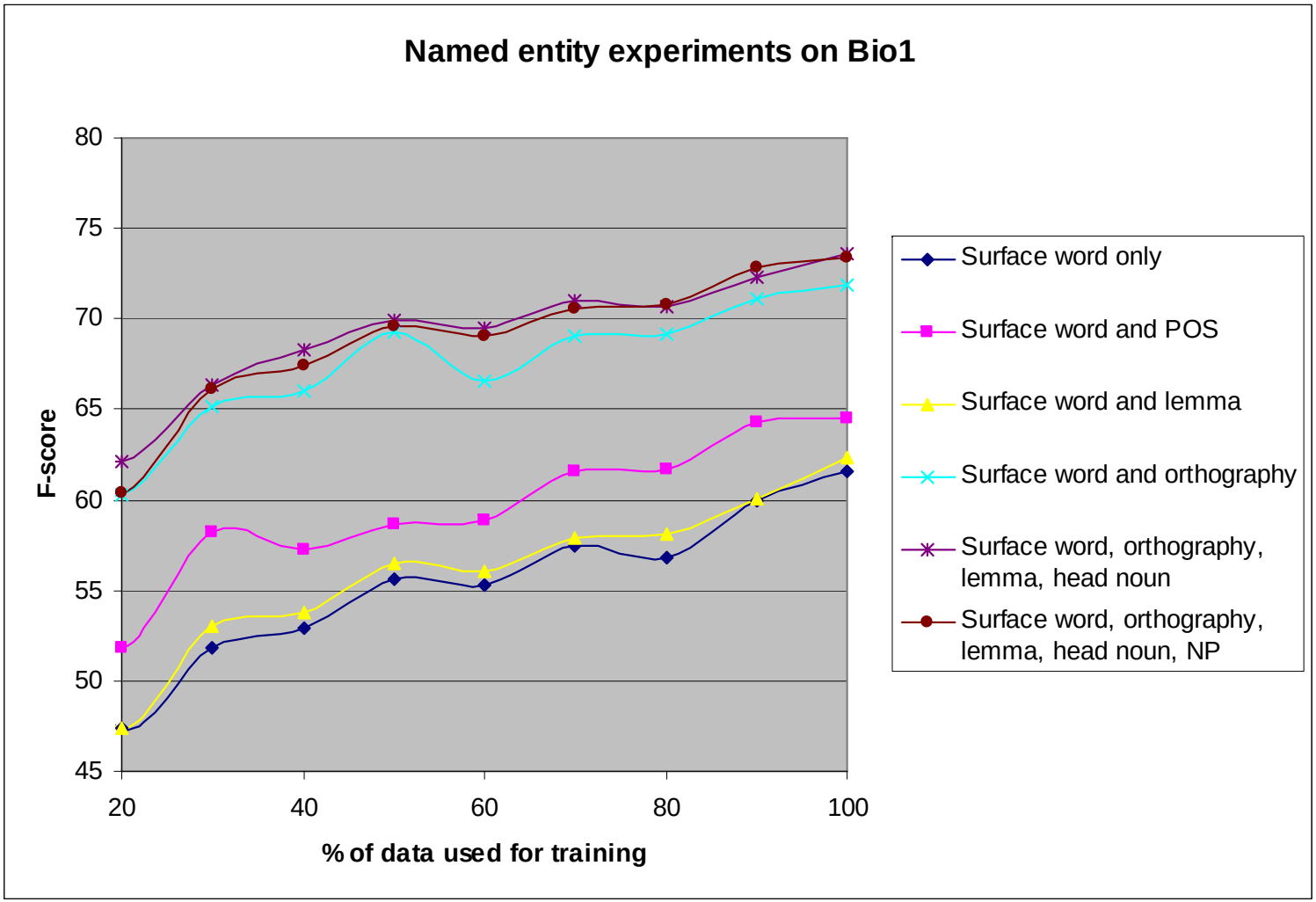
- Intuition
  - We've seen:
  - $[\text{LMP\_cap -\_hyp 1\_dig}]_{\text{protein}}$
  - Hypothesize that:
  - $[\text{AP\_cap -\_hyp 2\_dig}]_{\text{protein}}$

Results [5]; Words, orthography, lemma and head noun

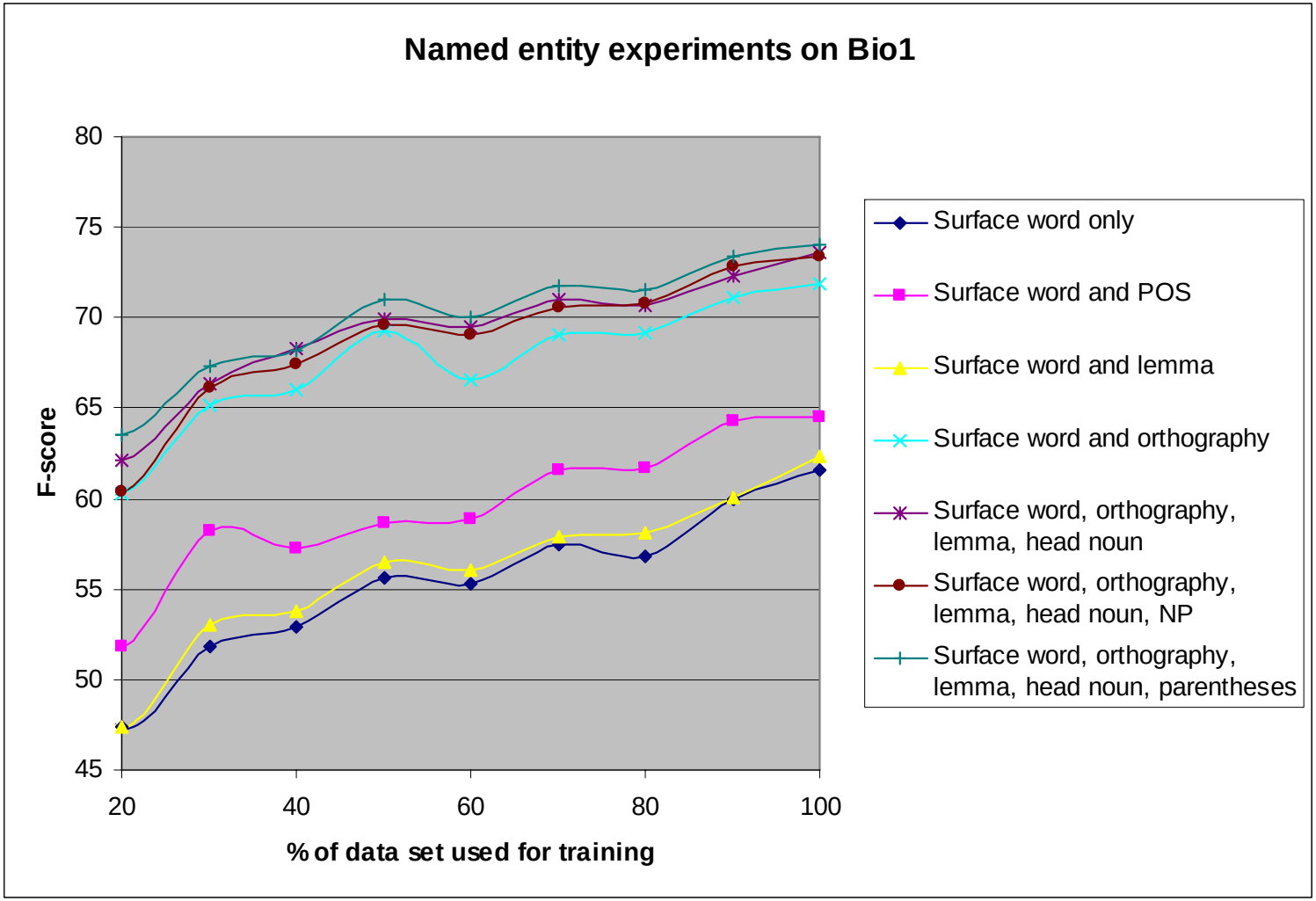


- Intuition
  - We've seen:
  - [T_cell cells_cell]$_{\text{cell type}}$
  - Hypothesize that:
  - [breast_cell carcinoma_cell cells_cell]$_{\text{cell type}}$

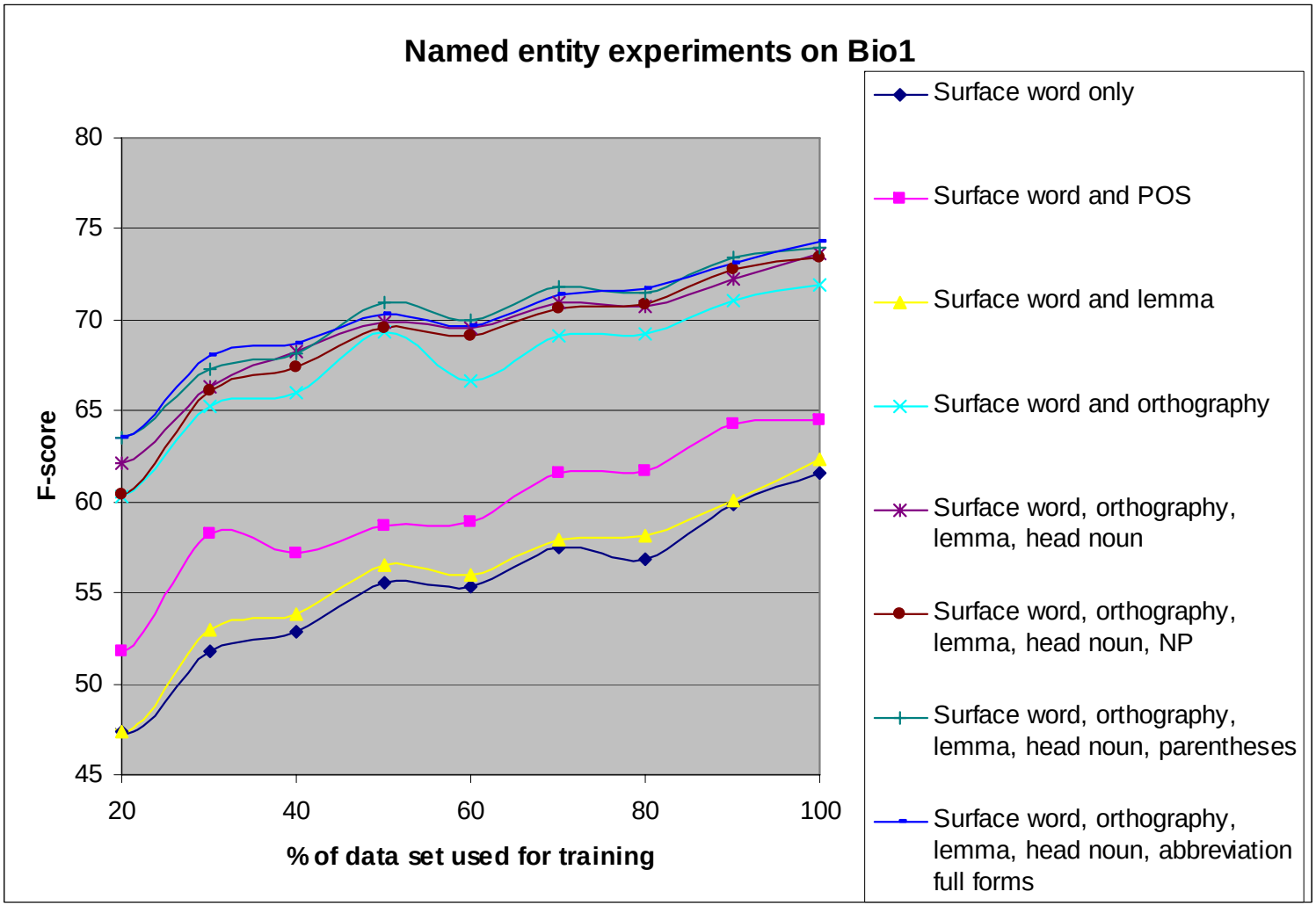
Results [6]: Words, orthography, lemma, head noun, noun phrase



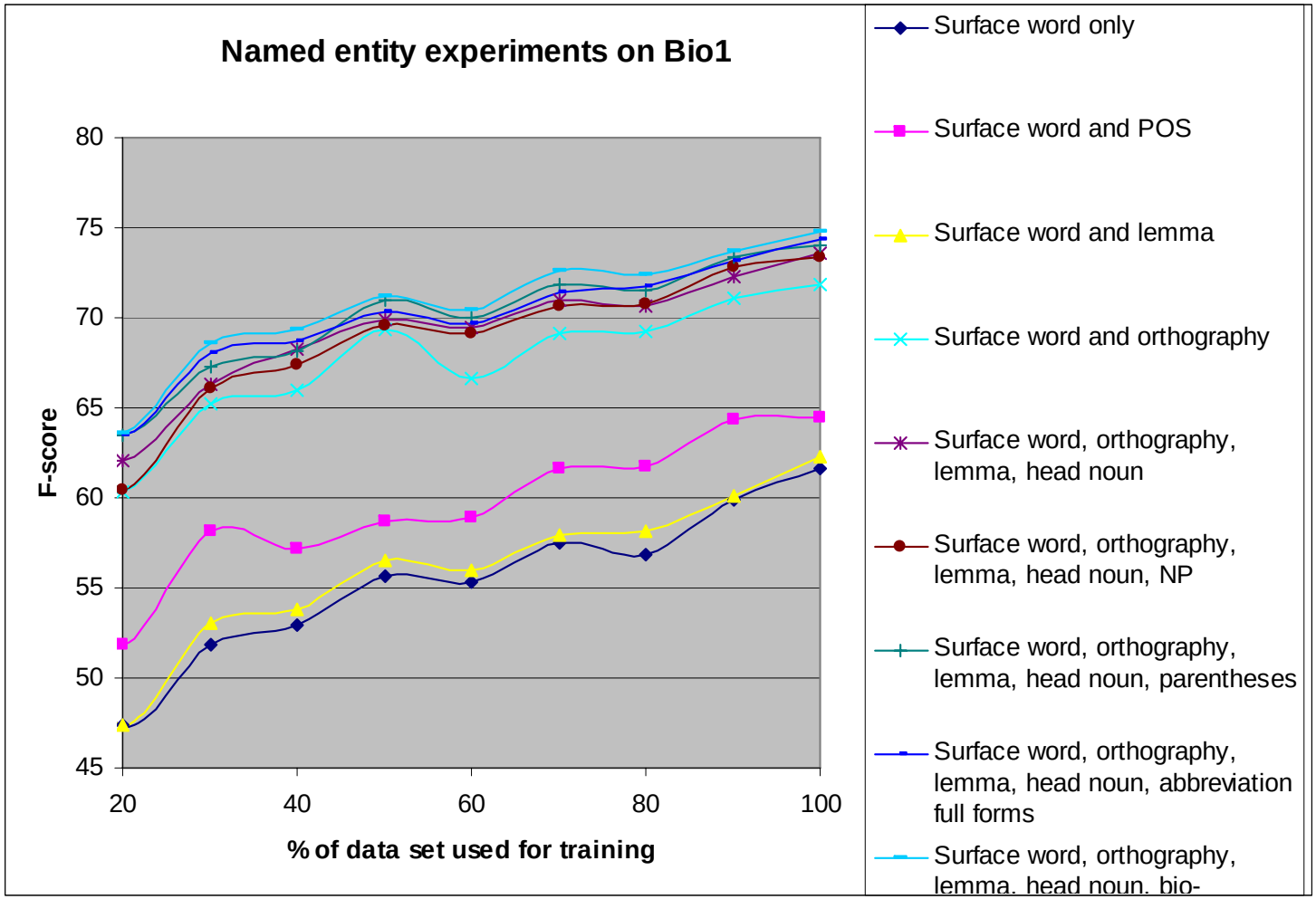
Results [7]: Words, othography, lemma, head noun, parentheses



Results [8]: Words, orthography, lemma, head noun, abbreviations

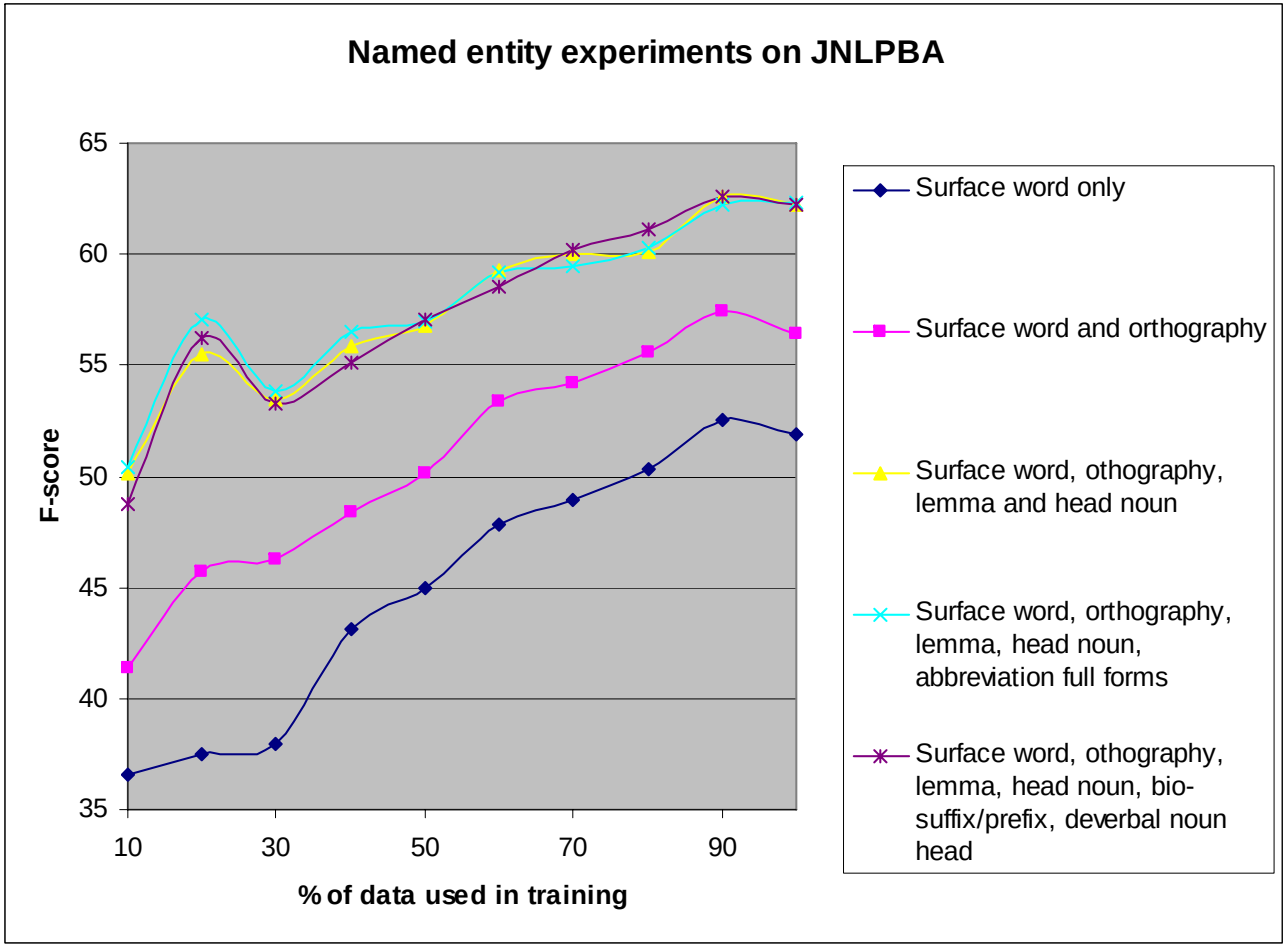


Results [9]: Words, orthography, lemma, head noun, bio-suffixes/prefixes



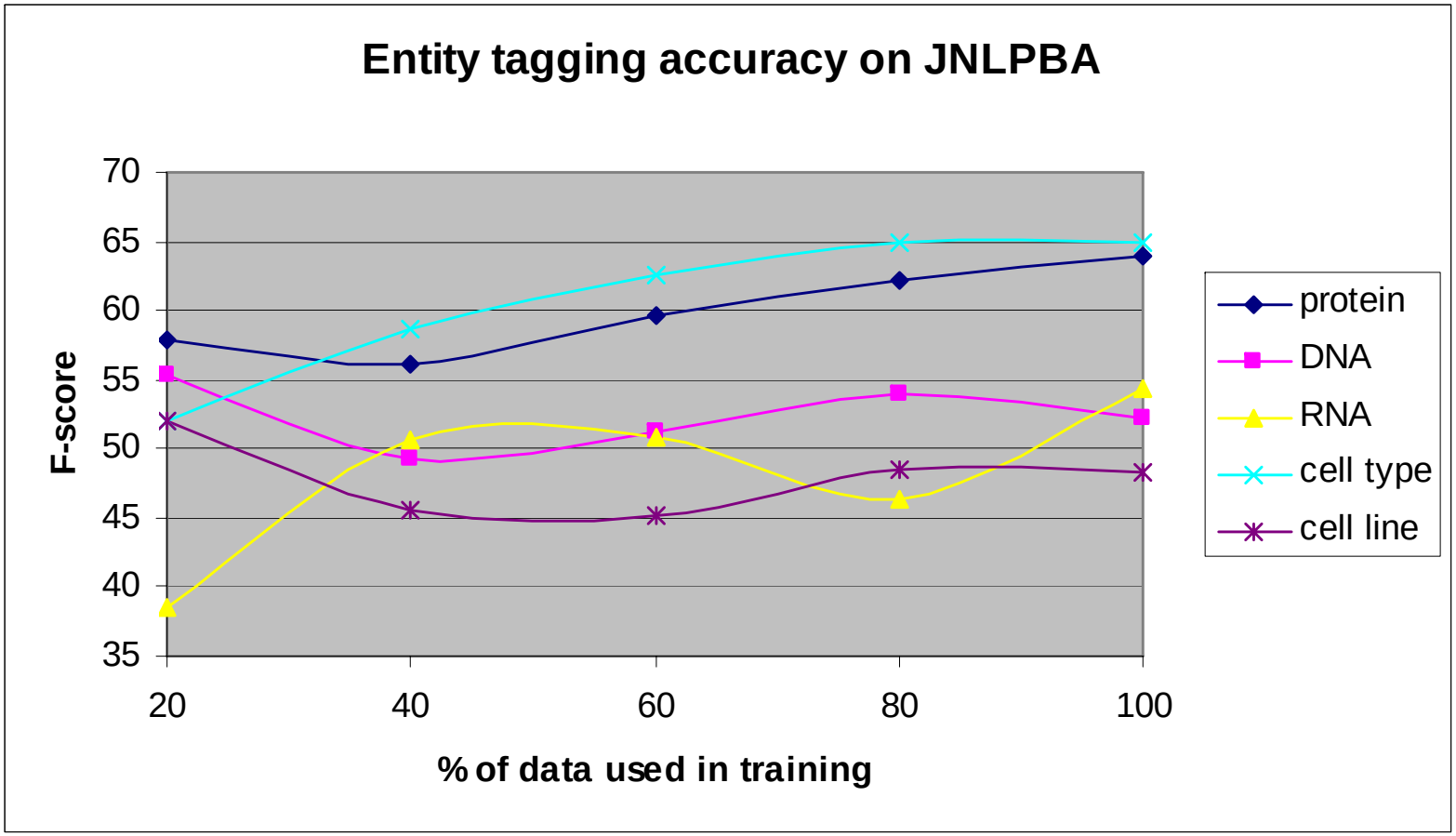
Influence of data set size on models

- Results for selected models on JNLPBA



Influence of data set size on classes

- Results for the best model on JNLPBA

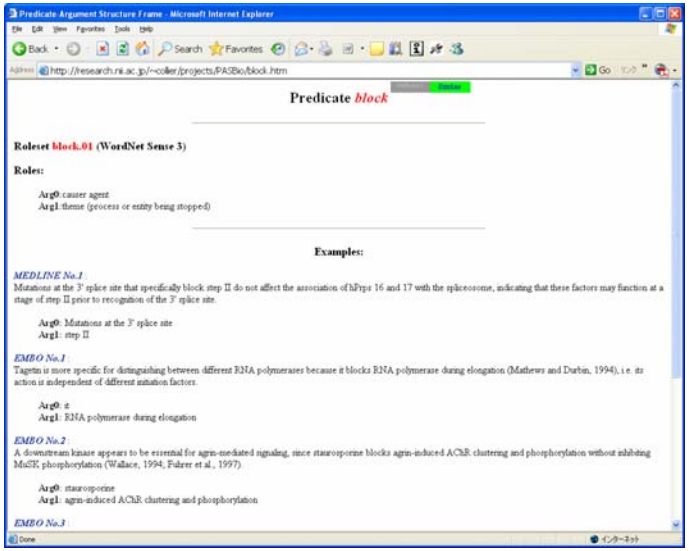


Confusion matrix for JNLPBA

#	Value	1	2	3	4	5	6	7	8	9	10	11	error
1	O	77.9	0.8	0.2	0.4	0.1	0.2	0.0	0.0	0.0	0.0	0.0	2.7
2	B-protein	1.3	3.3	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	1.6
3	I-protein	1.5	0.3	2.8	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	2.0
4	B-cell type	0.4	0.01	0.0	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.6
5	I-cell type	0.6	0.0	0.1	0.2	2.1	0.0	0.0	0.0	0.1	0.0	0.0	1.0
6	B-DNA	0.3	0.2	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.5
7	I-DNA	0.6	0.1	0.2	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.9
8	B-cell line	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.3
9	I-cell line	0.3	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.4	0.0	0.0	0.6
10	B-RNA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1
11	I-RNA	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1
	error	5.4	1.5	1.3	0.6	0.7	0.3	0.4	0.1	0.2	0.0	0.1	10.5

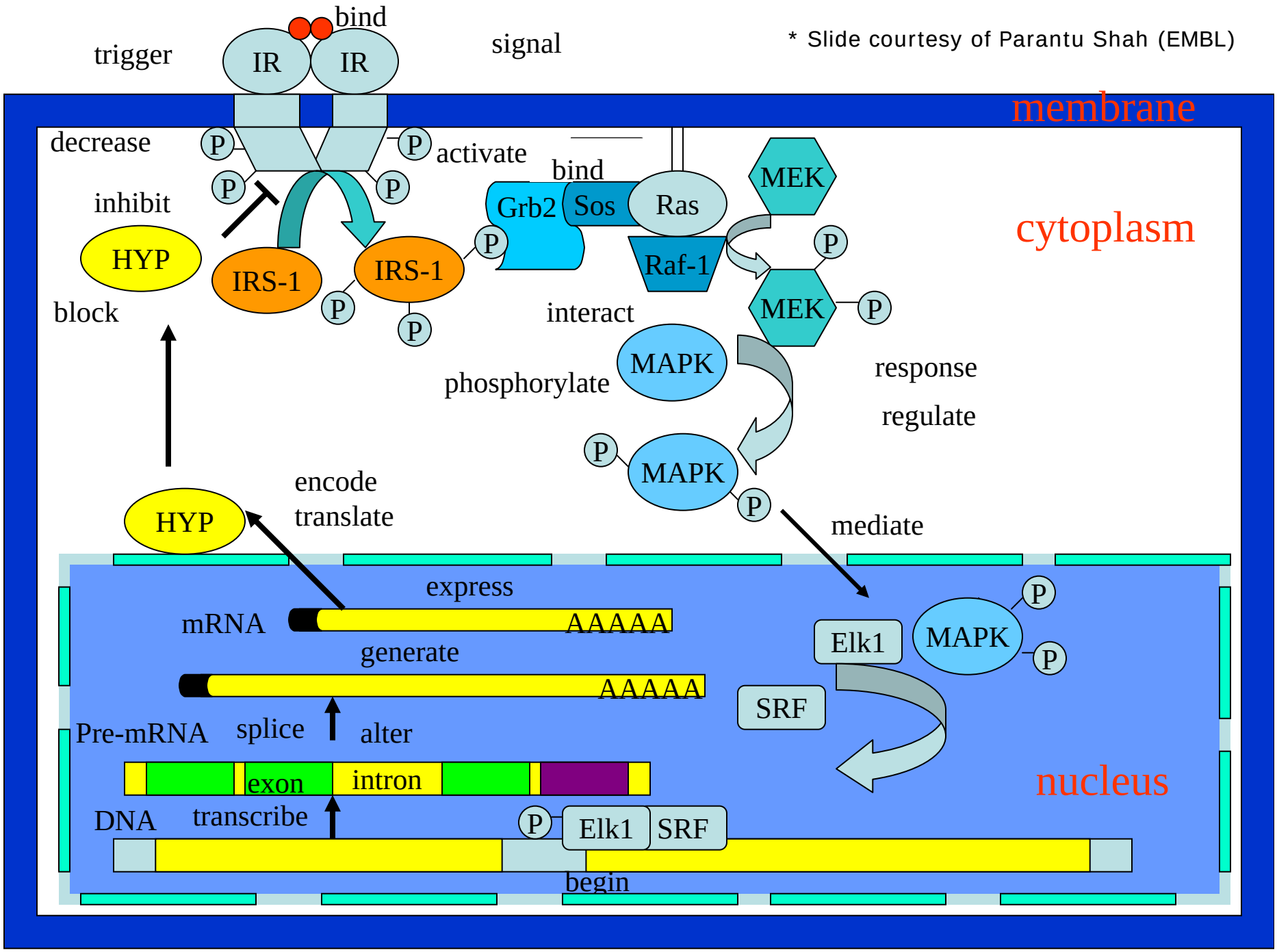
Analysing PAS frames in biology: PASBio

- Extend Propbank [Kingsbury and Palmer, 2002]
- Collect a corpus of domain texts
- Identify the major verbs (predicates) that indicate events
- Extract example sentences
- Analyse verb senses and argument roles with domain experts
- See if this fits with PropBank's existing frames
- If not then add a new frame and annotate selected sentences
- Perform machine learning to automate annotation

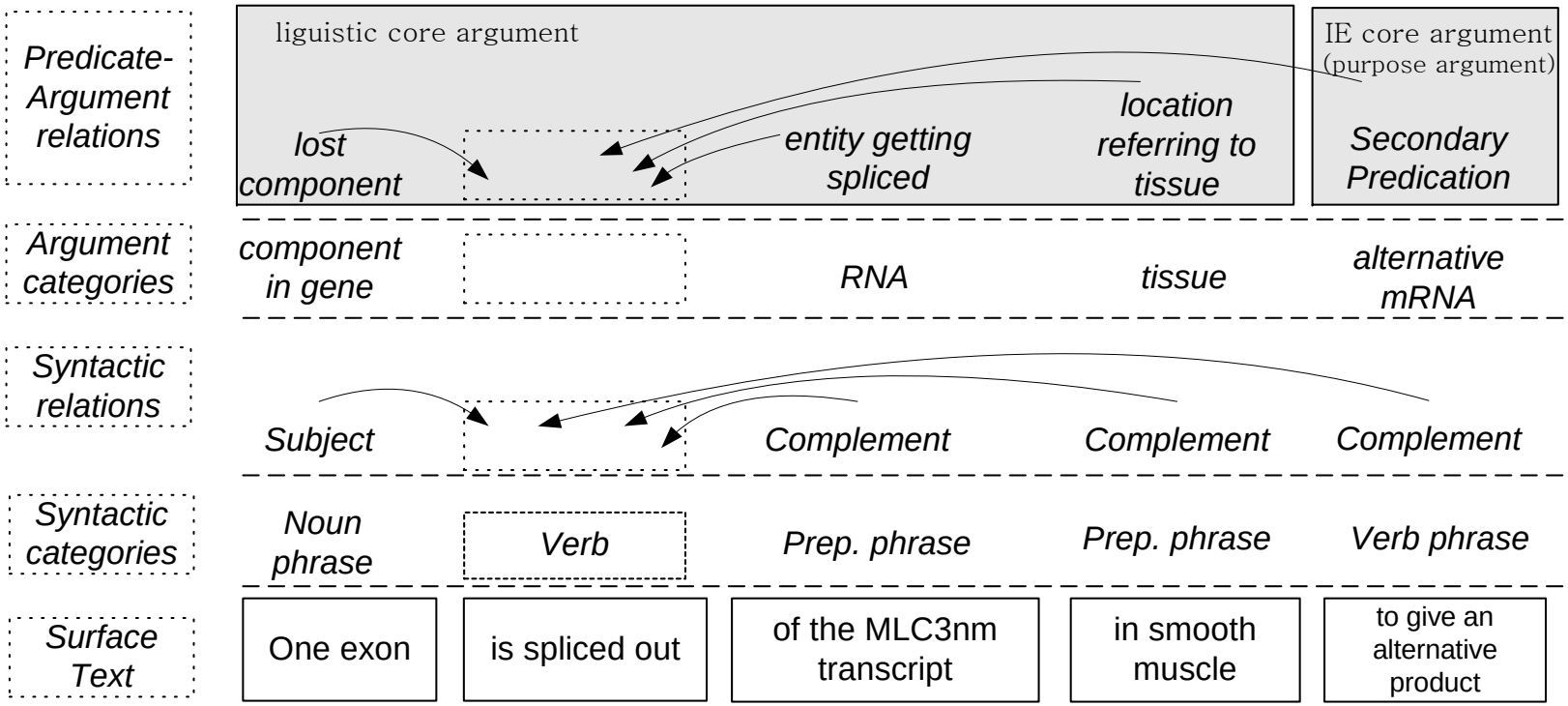


http://research.nii.ac.jp/~collier/projects/PASBio/index.html

Wattarujeekrit, T., Shah, P. and Collier, N. (2004), "PASBio: predicate-argument structures for event extraction in molecular biology", in BMC Bioinformatics, 5:155.


trigger
bind
IR
IR
signal
* Slide courtesy of Parantu Shah (EMBL)
membrane
decrease
P
P
activate
P
P
inhibit
HYP
block
IRS-1
IRS-1
P
P
P
bind
Grb2
Sos
Ras
Raf-1
MEK
MEK
P
P
cytoplasm
interact
MAPK
phosphorylate
response
regulate
P
MAPK
P
encode
translate
HYP
mediate
express
mRNA
AAAAA
generate
AAAAA
Pre-mRNA
splice
alter
exon
intron
DNA
transcribe
P
Elk1
SRF
begin
Elk1
MAPK
P
P
SRF
nucleus

Using predicate argument information to constrain types



Wattarujeekrit, T. and Collier, N. (2005), in proceedings of the Eighth International Conference on Discovery Science, Singapore.

Scores for selected predicates on JNLPBA

Predicate	Lexical model	Lexical model + dependency path + voice + head pair + trans/intrans	Improvement
Encode	56.6	57.6	+1.0
Recognize	47.2	49.4	+2.2
Block	51.2	52.0	+0.8
Lead	57.0	57.5	+0.5

Discussion and future work

Developments in JNLPBA

- State-of-the-art systems have taken the feature sets into new areas:
  - Use of character n-grams for affix features
  - Use of gazetteers (derived from LocusLink, GO etc.)
  - Use of syntactic information
  - Use of external resources (BNC, Google search)
  - Context holding mechansims (previously predicted entities)

System	Features	F-score
Zho	af,or,gn,gz,po,tr,ab,ca	72.6
Fin	lx,af,sh,gz,po,sy,ab,do,pa	70.1
Set	lx,af,or,sh,gz,tr	69.8
Son	af,or,po,np	66.3
Zha	lx	64.8
Rös	af,or,gn,ln	64.0
Par	Af,or,sh,gn,wv,po,np,tr	63.0
Lee	Af,po	49.1

Final thought [1]

- State of the art still seems far away from human performance
  - Maybe 80 F-score is good enough for practical applications? Must be led by biologists needs.
- But what *is* human performance?
  - About 97% for MUC-7 on news data
  - We have some evidence, e.g. 87% (Hirschman, 2003), 89% (Demetrious and Gaizauskas, 2003)
  - .. but not really enough – people are too busy doing NE to consider the task itself!
  - Need inter-annotator agreement scores and intra-annotator agreement scores

Final thought [2]

- But what is the ‘right’ level of knowledge?
- A study of IAA or NE should also consider what levels of knowledge the annotators use to make their decisions:
  - Sentential
  - Document
  - World knowledge
  - Guess

Final thought [3]

- What kinds of ontologies are appropriate for annotation of text spans?
- BioCreative (2004) 1b normalization of gene names showed one good way
    - Mapping entities to some conceptual definition in an ontology
    - Combines named entity with coreference resolution on real world ontologies
    - But seems to add a level of complexity
- As a community we need to decide on a consensus for the way forward – traditional MUC-style NER or ontology class mapping or a combination of both

Conclusion

- Biomedical NER has been successful
  - A step forward into defining semantics in domain texts
  - Resources were created and re-used, models adapted, tools deployed – but not nearly enough deployment yet
  - Started us on the track to disciplined methodologies and open evaluations
  - Insights into the nature of terminology in the domain
- Not far enough yet?
  - 80 F-score seems to be the upper limit, but why? Is 80 F-score enough? Is it the task definition, the data or the knowledge-level?
- Cross domain comparisons
  - No formal way yet to compare difficulties across domains (e.g. news vs biology, EMBOJ vs Nature, different subsets of MEDLINE)

Acknowledgements

- Terminology annotation
  - Koichi TAKEUCHI (Okayama U.)
- Coreference annotation
  - Ai KAWAZOE (NII)
  - Asanobu KITAMOTO (NII)
- Predicate-argument annotation
  - Tuangthong WATTARUJEEKRIT (NII)
  - Parantu SHAH (EMBL)
- Rhetorical structure annotation
  - Yoko MIZUTA (NII)
  - Anthony MULLEN (Tsuda, U.)
- Verb semantics
  - Anna KORHONEN (Cambridge U.)
- Bio-domain knowledge
  - Shoko KAWAMOTO (NAIST, NII)
  - Asao FUJIYAMA (NII, RIKEN)
- JNLPBA 2004 shared task
  - Jin-Dong KIM (U. Tokyo)
  - Tomoko OHTA (U. Tokyo)
  - Yoshimasa TSURUOKA (U. Tokyo)
  - Yuka TATEISHI (U. Tokyo)
- Funding
  - NII, JSPS

Recent Publications [1]

- Collier, N., Nazarenko, A., Baud, R. and Ruch, P. (2005) “Recent advances in natural language processing for biomedical applications”, in vol. 74, no. 11, International Journal of Medical Informatics, Elsevier (in press).
- Mizuta, Y., Korhonen, A., Mullen, A. and Collier, N. (2005), “Zone analysis in biology articles as a basis for information extraction”, in vol. 74, no. 11, International Journal of Medical Informatics, Elsevier (in press).
- Yacov Kogan, Nigel Collier, Serguei Pakhomov and Michael Krauthammer (2005), "Towards Semantic Role Labeling & IE in the Medical Literature", in proceedings of the American Medical Informatics Association annual symposium, Washington DC, USA, October 22-26 (in press).
- Mullen, T., Mizuta, Y. and Collier, N. (2005), "Classifying rhetorical zones in biology texts with a support vector machine", in vol. 7, no. 1, SIGKDD Explorations.
- Takeuchi, K. and Collier, N. (2005), "Bio-medical entity extraction using support vector machines", in vol. 33, no. 2, Artificial Intelligence in Medicine, Elsevier, pp. 125-137, DOI information: 10.1016/j.artmed.2004.07.019.
- Wattarujeekrit, T., Shah, P. and Collier, N. (2004), “PASBio: predicate-argument structures for event extraction in molecular biology”, in BMC Bioinformatics, 5:155.
- Collier, N. and Takeuchi, K. (2004), “Comparison of character-level and part of speech features for name recognition in bio-medical texts” in vol. 37, no. 6, Journal of Biomedical Informatics, Elsevier, December, pp. 423-435

Recent Publications [2]

- Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y. and Collier, N. (2004), "Introduction to the Bio-Entity Recognition Task at JNLPBA", in proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 28-29 August, Geneva, Switzerland.
- Mizuta, Y. and Collier, N. (2004), "Zone Identification in Biology Articles as a Basis for Information Extraction", in proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications held at COLING'2004, 28-29 August, Geneva, Switzerland.
- Mizuta, Y. and Collier, N. (2004), "An Annotation Scheme for Rhetorical Analysis of Biology Articles", in proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004), 26-28 May, Lisbon, Portugal.
- Kawazoe, A., Kitamoto, A. and Collier, N. (2004), "Annotation of Coreference Relations among Linguistic Expressions and Images in Biomedical Articles", in proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004), 26-28 May, Lisbon, Portugal, pp. 529-532.
- Kawazoe, A. and Collier, N. (2003), "Open Ontology Forge: A Tool for Ontology Creation and Text Annotation in a Biomedical Domain", in proceedings of the 14th Conference on Genome Informatics, 14-17 December, Yokohama, Japan.
- Collier, N., Takeuchi, K., Kawazoe, A., Mullen, A. and Wattarujeekrit, T. (2003), "A Framework for Integrating Deep and Shallow Semantic Structures in Text Mining", in proceedings of the Seventh International Conference on Knowledge-based Intelligent Information and Engineering Systems (KES'2003), University of Oxford, UK, 3-5 September. (Also published by Springer-Verlag as a book part).

Recent Publications [3]

- Collier, N., Takeuchi, K. and Kawazoe, A. (2003), “Open Ontology Forge: An Environment for Text Mining in a Semantic Web World”, in proceedings of the Semantic Web Foundations and Application Technologies Workshop (SWFAT), Nara, Japan, March 12th, 2003.
- Kawazoe, A. and Collier, N. (2003), “An Ontologically-motivated Annotation Scheme for Coreference”, in proceedings of the Semantic Web Foundations and Application Technologies Workshop (SWFAT), Nara, Japan, March 12th, 2003.
- Takeuchi, K. and Collier, N. (2002), “Use of Support Vector Machines in Extended Named Entity Recognition”, in proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002), Taipei, Taiwan, August.
- Collier, N., Takeuchi, K., Nobata, C., Fukumoto, J. and Ogata, N. (2002), “Progress on Multi-lingual Named Entity Annotation Guidelines using RDF(S)", in proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, 29th – 31st May, pp. 2074-2081.
- Collier, N. and Takeuchi, K., (2002), “PIA-Core: Semantic Annotation through Example-based Learning” in proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, 29th – 31st May, pp. 1611-1614.
- Collier, N., Nobata, C., and Tsujii, J. (2001), "Automatic acquisition and classification of molecular biology terminology using a tagged corpus", vol. 7, no. 2, Journal of Terminology, John Benjamins pp. 239-258.