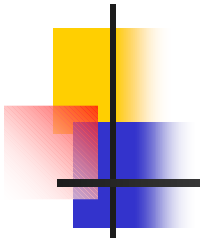
Machine Learning for Information Extraction from XML marked-up text on the Semantic Web

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

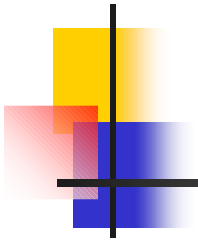
May 1st 2001

Semantic Web Workshop 2001 at WWW10



Talk summary

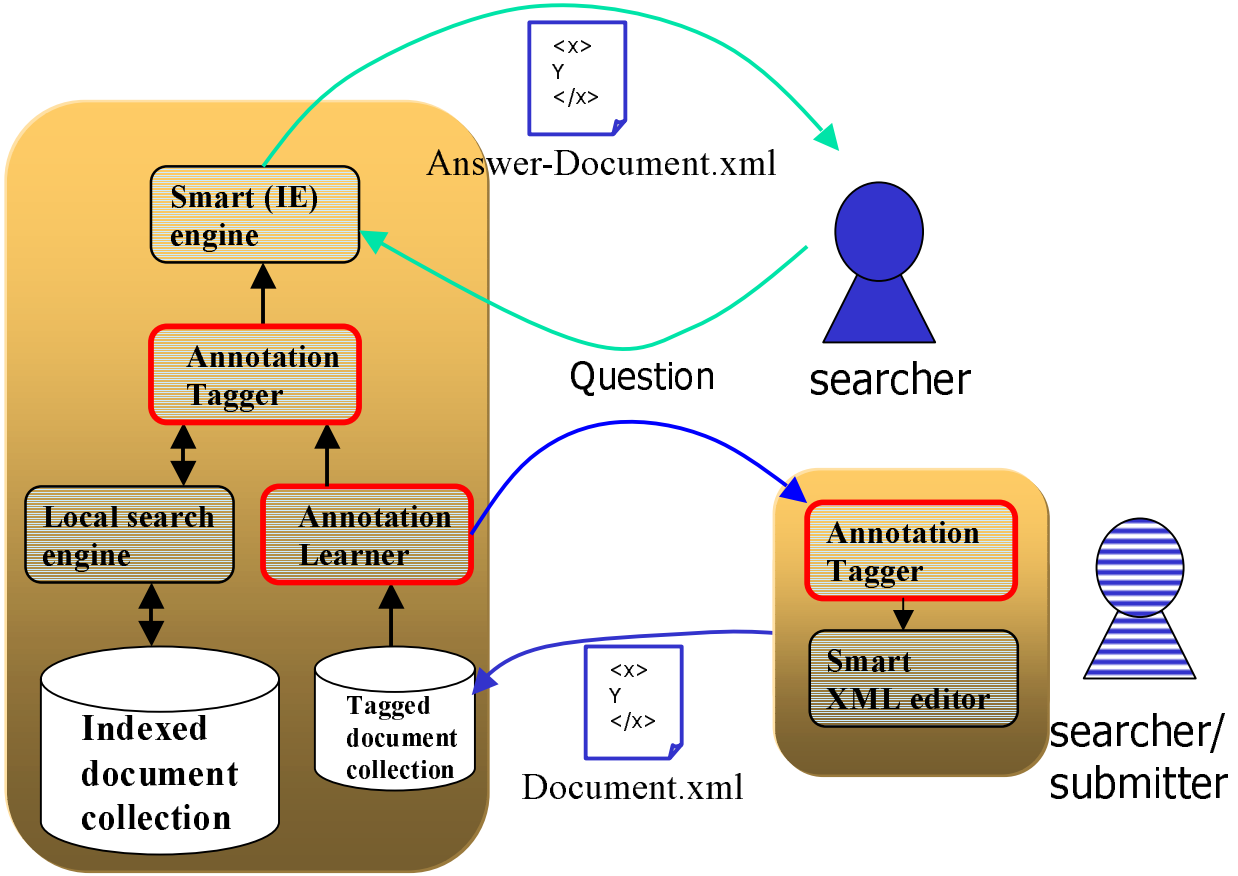
- Introduction
  - Motivation
  - System model
  - Tagged texts as the key to learning
- Test collections
- Method
- Results and Conclusion

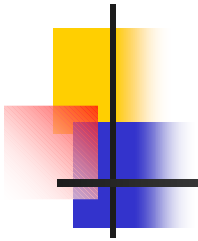


Introduction and motivation

- Final goal:
  - Smart documents and smart applications based on standardised content annotation schemes XML, RDF etc..
- Why is this a good thing?
  - Information access, building natural interfaces etc.
- The bottleneck:
  - Entering expert knowledge into (textual) documents
- Proposed solution:
  - Learning to annotate domain-based texts using examples.

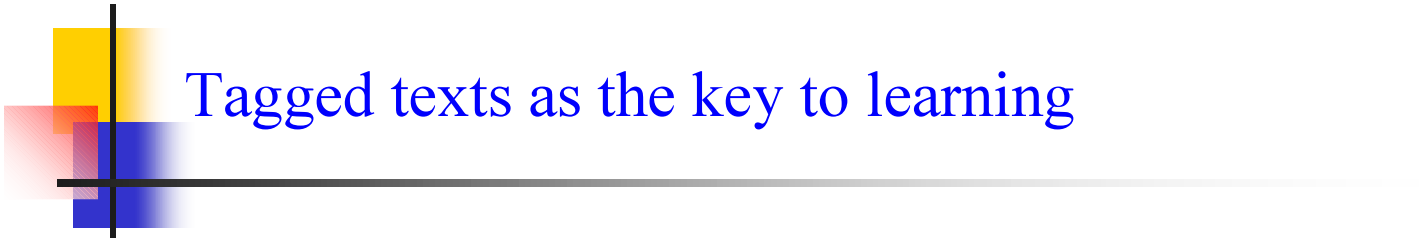
System model: PIA project at NII





System model

- Initial goals:
  - a pilot study to test machine learning technology in a technical domain as well as news.
  - explore the problems of tagging from a linguistic perspective.
  - Concentrate on terminology, i.e. identification & classification of terms
  - using examples to learn
- Next step goals:
  - Make use of higher level information contained in the DTD schema, attribute information etc. Define and use ontologies etc..



Tagged texts as the key to learning

•Example marked-up sentence for molecular-biology:

No <PROTEIN>STAT</PROTEIN> activity was detected in <SOURCE subtype=ct>TCR-stimulated lymphocytes</SOURCE>, indicating that the <PROTEIN>JAK</PROTEIN>/<PROTEIN>STAT</PROTEIN> pathway defined in this study constitutes an <PROTEIN>IL-2R</PROTEIN>- mediated signaling event which is not shared by the <PROTEIN>TCR</PROTEIN>.

Challenges of name-finding in a technical domain

- Inconsistent naming conventions

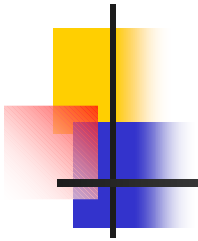
  e.g. IL-2, IL2, Interleukin 2, Interleukin-2, Il-2

- Wide-spread synonymy

  Many synonyms in wide usage, e.g. PKB and Akt

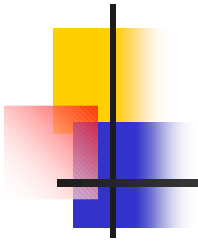
- Open, growing vocabulary for many classes

- Cross-over of names between classes depending on context



HMM models

- Advantages
  - can consider language modeling within a well-known and understood mathematical framework
  - although the n-1 assumption is naïve, it works well in practice
- Disadvantages
  - the model ignores long distance and structural dependencies
  - the model suffers from fragmentation of probability distribution (i.e. data sparseness)



Model specification

- Formal generative model

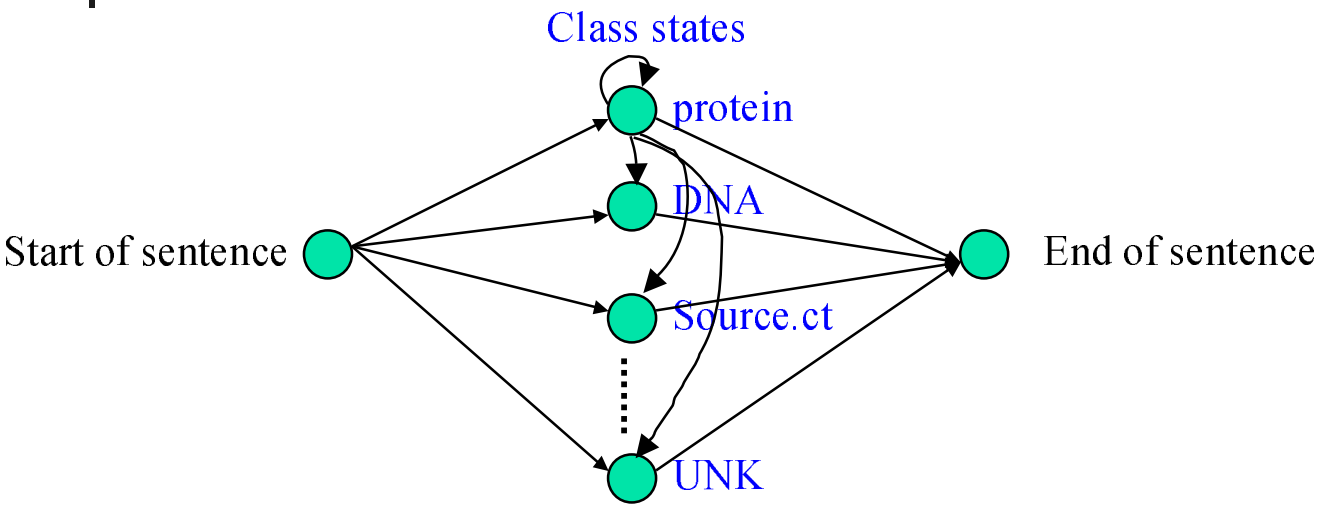
$$\Pr(NC \mid W) = \frac{\Pr(W, NC)}{\Pr(W)}$$

NC a sequence of name classes

W, a given sequence of words

Since Pr(W) can be considered to be constant we aim to maximize Pr(W,NC).

Model's intuition



Example:

Activation of JAK kinases and STAT proteins in human T lymphocytes .

UNK UNK PROTEIN PROTEIN UNK PROTEIN PROTEIN UNK SOURCE.ct SOURCE.ct SOURCE.ct UNK

Underlying process:

Interpolating HMM model specification

- We need two probability distributions
  - (1) for the first word and name class in a sequence
  - (2) for all other words and name classes
- Let (1) be,

$$\sigma_0 \Pr(NC_{first} | < W_{first}, F_{first} >) +$$
$$\sigma_1 \Pr(NC_{first} | < \_, F_{first} >) +$$
$$\sigma_2 \Pr(NC_{first})$$
$$for$$
$$\sum \sigma_i = 1.0,$$
$$\sigma_0 \geq \sigma_1 \geq \sigma_2$$

□

σ_x	empirically determined constant
NC_{first}	first name class (state) in the sequence
W_{first}	first word in observed emission
F_{first}	first feature belonging to first word

Interpolating HMM model specification

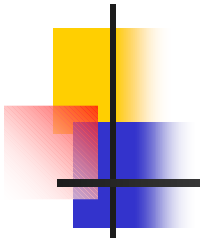
- Let (2) be,

$$\lambda_0 \Pr(NC_t \mid < W_t, F_t >, < W_{t-1}, F_{t-1} >, NC_{t-1}) +$$
$$\lambda_1 \Pr(NC_t \mid < \_, F_t >, < W_{t-1}, F_{t-1} >, NC_{t-1}) +$$
$$\lambda_2 \Pr(NC_t \mid < W_t, F_t >, < \_, F_{t-1} >, NC_{t-1}) +$$
$$\lambda_3 \Pr(NC_t \mid < \_, F_t >, < \_, F_{t-1} >, NC_{t-1}) +$$
$$\lambda_4 \Pr(NC_t \mid NC_{t-1}) +$$
$$\lambda_5 \Pr(NC_t)$$
$$for$$
$$\sum \lambda_i = 1.0,$$
$$\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_5$$

□

λx	empirically determined constant
NC_t	next name class (state) in the sequence
W_t	next word in observed emission
F_t	next feature belonging to first word

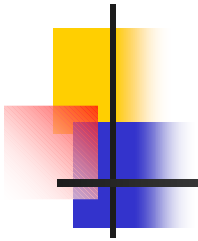
- The optimal path is recovered using the Viterbi algorithm



Interpolating HMM model specification

Character features:

Code	Feature	Example
dig	DigitNumber	15
sin	SingleCapital	M
grk	GreekLetter	alpha
cad	CapsAndDigits	I2
cap	AtLeastTwoCap	RaIGDS
lad	LettersAndDigit	i12
fst	FirstWord	(first word in sentence)
ini	InitCap	Interleukin
lcp	LowerCaps	kappaB
low	LowerCase	kinases
hyp	Hyhon	- '
opp	OpenParenthese	(
clp	CloseParenthes	)
fsp	FullStop	.
cma	Comma	,
pct	Percent	%
osq	OpenSquareBra	[
csq	CloseSquareBra	]
cln	Colon	:
scn	SemiColon	;
det	Determiner	the
con	Conjunction	and
oth	Other	*+#@

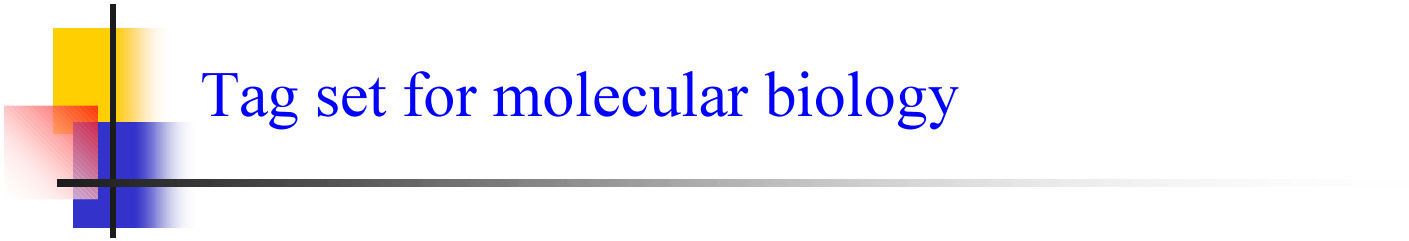


Experiments (molecular biology)

- Interpolating HMM (NEHMM)

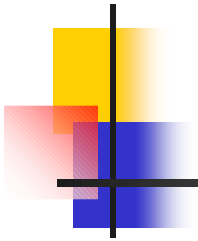
- Domain of biochemistry: *human+blood cell+transcription factor*

- Corpus:
  - 100 MEDLINE abstracts -
    - 80 for training, 20 for testing with 5-fold cross-validation
  - Tagged by domain expert
  - Developed at the Tsujii laboratory (U. Tokyo)
- Ontology:
  - A simple taxonomy that forbid term class overlapping
  - based on substance characteristics (rather than e.g. role)



Tag set for molecular biology

Class	#	Example	Description
PROTEIN	2125	JAK kinase	proteins, protein groups, families, complexes and substructures.
DNA	358	IL-2 promoter	DNAs, DNA groups, regions and genes
RNA	30	TAR	RNAs, RNA groups, regions and genes
SOURCE.cl	93	leukemic T cell line Kit225	cell line
SOURCE.ct	417	human T lymphocytes	cell type
SOURCE.mo	21	Schizosacchar-omyces pombe	mono-organism
SOURCE.mu	64	mice	multi-organism
SOURCE.vi	90	HIV-1	viruses
SOURCE.sl	77	membrane	sub-location
SOURCE.ti	37	central nervous system	tissue
UNK	-	tyrosine phosphorylation	background words

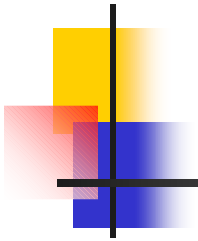


Experiments (news)

- Interpolating HMM (NEHMM)

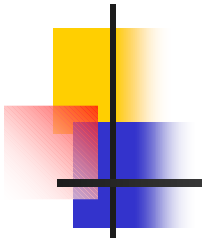
- Domain of news: MUC-6 dry run and formal run test set

- Corpus:
  - 60 news texts -
    - 50 for training, 10 for testing with 6-fold cross-validation
- Ontology:
  - No explicit ontology. MUC-6 tagging guidelines.



Tag set for news

Class	#	Example	Description
ORGANISATION	1783	Harvard Law School	names of organisations
PERSON	838	Washington	names of people
LOCATION	390	Houston	names of places, countries etc.
DATE	542	1970s	date expressions
TIME	3	midnight	time expressions
MONEY	423	$ 10 million	money expressions
PERCENT	108	2.5%	percentage expressions
UNK	-	start-up costs	background words

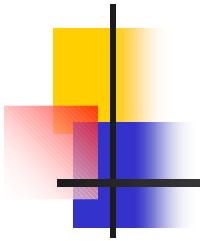


Results for news tests - comparison with molecular biology tests

System	News	Biology
HMM (w/Unity)	78.4	75.0
HMM (w/o Unity)	74.2	73.1

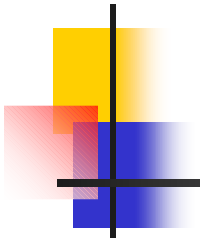
Table 2: F-score all class averages for news and molecular biology test sets

F-score = (2 x Precision x Recall) / (Precision + Recall)



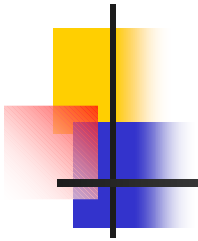
Analysis

- Classification was far easier than identification due to linguistic structures such as:
  - Coordination, e.g. c-rel and v-rel (proto) oncogenes
  - Apposition, e.g. The transcription factor NF-Kappa B..
  - Abbreviation, e.g. the Interleukin-2 (IL-2) promoter..



Analysis

- Ways forward:
  1. Use some other identification method than HMM?
  2. We estimate that the training texts are no more than 95% consistent between human-taggers - improve the consistency of tagging with better guidelines?
  3. Incorporate nested tagging to model term-internal dependencies? Or a domain independent dependency analyser.



Conclusion

1. The HMM performed quite well overall considering training data size.

2. Local context and small feature set limitations of the HMM need to be overcome in future models for complex local linguistic structures.

3. The model needs to make use of element type name relations such as combination relations and element attributes held inside the DTD as well as integrating ontological knowledge held e.g. in RDF(S).