

A Priority Model for Name Classification

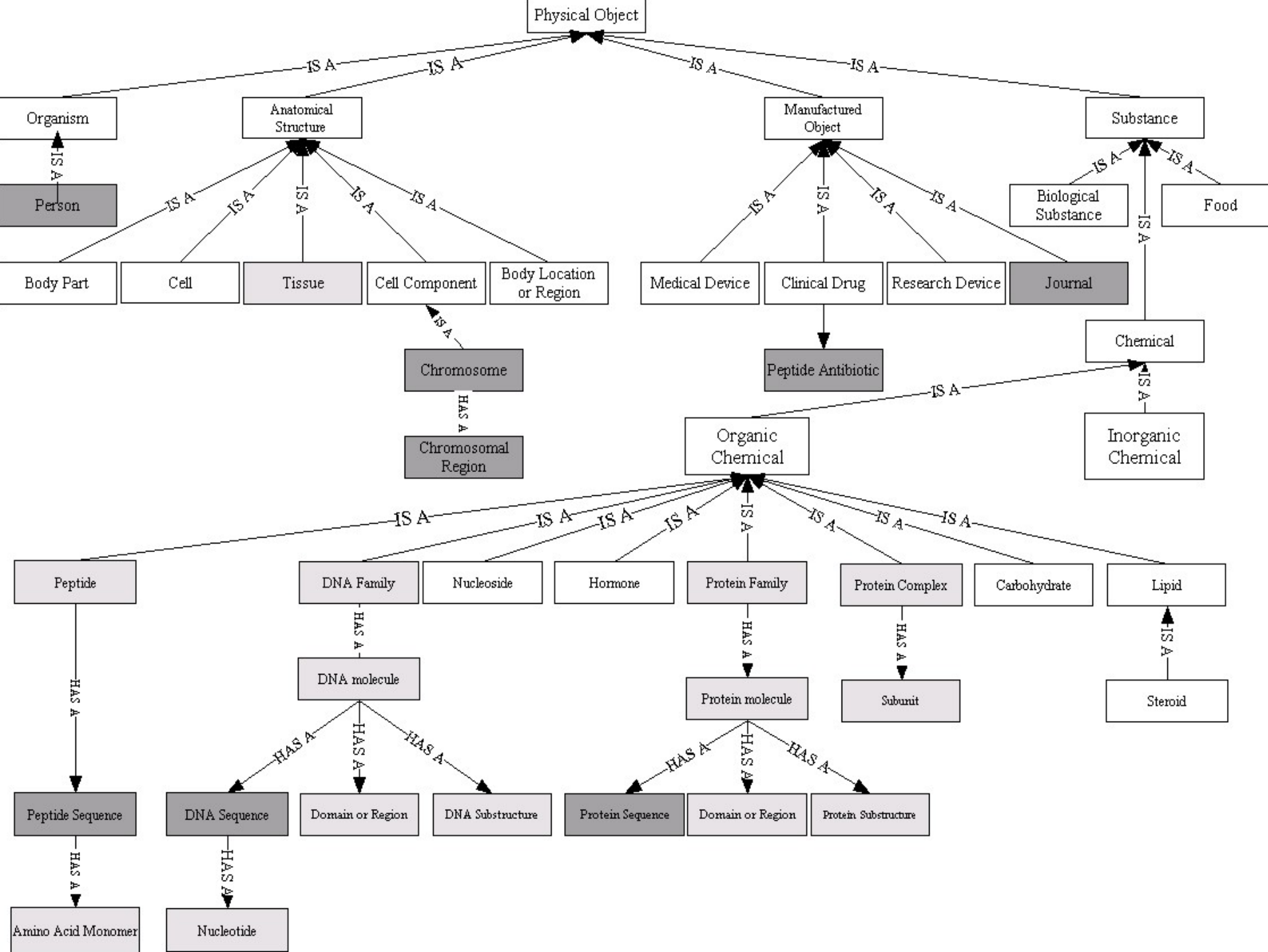
By John Wilbur and Lorrie Tanabe,
NCBI, NLM, NIH

Named Entity Recognition

- Textual Demarcation
- Classification
- Normalization
- Interplay

Classification

- Semantic network
- Borrowed from NLM, Genia, and some additions of our own
- SemCat Database



SemCat Sources

- UMLS, GENIA, UniProt, GO
- Entrez Gene, ProtScan, ChemID
- NCBI Taxonomy, Brown corpus
- Wall Street Journal Corpus
- WormBase, FlyBase
- Candida Genome Database
- Saccharomyces cerevisiae Database

SemCat Statistics

- 77 semantic categories
- 5.11 million names (some repeated)

Semantic Type	Sources	Total
CHEMICAL	UMLS, ChemId	1,246,237
PERSON	UMLS, NCBI author lists	1,118,774
DNA MOLECULE	GENIA, GO, WWW, Patterns	819,954
PROTEIN MOLECULE	UMLS, GO, ProtScan, Patterns	545,492
ORGANISM	NCBI Taxonomy	239,873
DISEASE/SYNDROME	UMLS	161,672
THERAPEUTIC	UMLS, Patterns	127,088
BODY PART	GENIA, UMLS, WWW	96,449
COMMON WORDS	Brown, Wall Street Journal	91,655
INJURY/POISONING	UMLS	84,602
MEDICAL DEVICE	UMLS, WWW	80,498
FINDING	UMLS	75,806
NEOPLASTIC	UMLS	45,607
CLINICAL DRUG	UMLS, ChemId, WWW	34,643
CELL	GENIA, NCI, ATCC, WWW	26,335

SemCat Classification

- Binary task: Gene/Protein names vs. Other names
- Random Train/Test Divisions 90%/10%
- Repeated 3 times

Classification Methods

- Language model
- Probabilistic Context Free Grammars
- Priority model

Language Model

- Bigram
- Witten-Bell Smoothing

PCFG-3

- $CATP \rightarrow CATP\ CATP$
- $CATP \rightarrow CATP\ postCATP$
- $CATP \rightarrow preCATP\ CATP$

PCFG-8

- $CATP \rightarrow CATP\ CATP$
- $CATP \rightarrow CATP\ postCATP$
- $CATP \rightarrow preCATP\ CATP$
- $CATP \rightarrow NotCATP\ CATP$
- $NotCATP \rightarrow NotCATP\ NotCATP$
- $NotCATP \rightarrow NotCATP\ postNotCATP$
- $NotCATP \rightarrow preNotCATP\ NotCATP$
- $NotCATP \rightarrow CATP\ NotCATP$

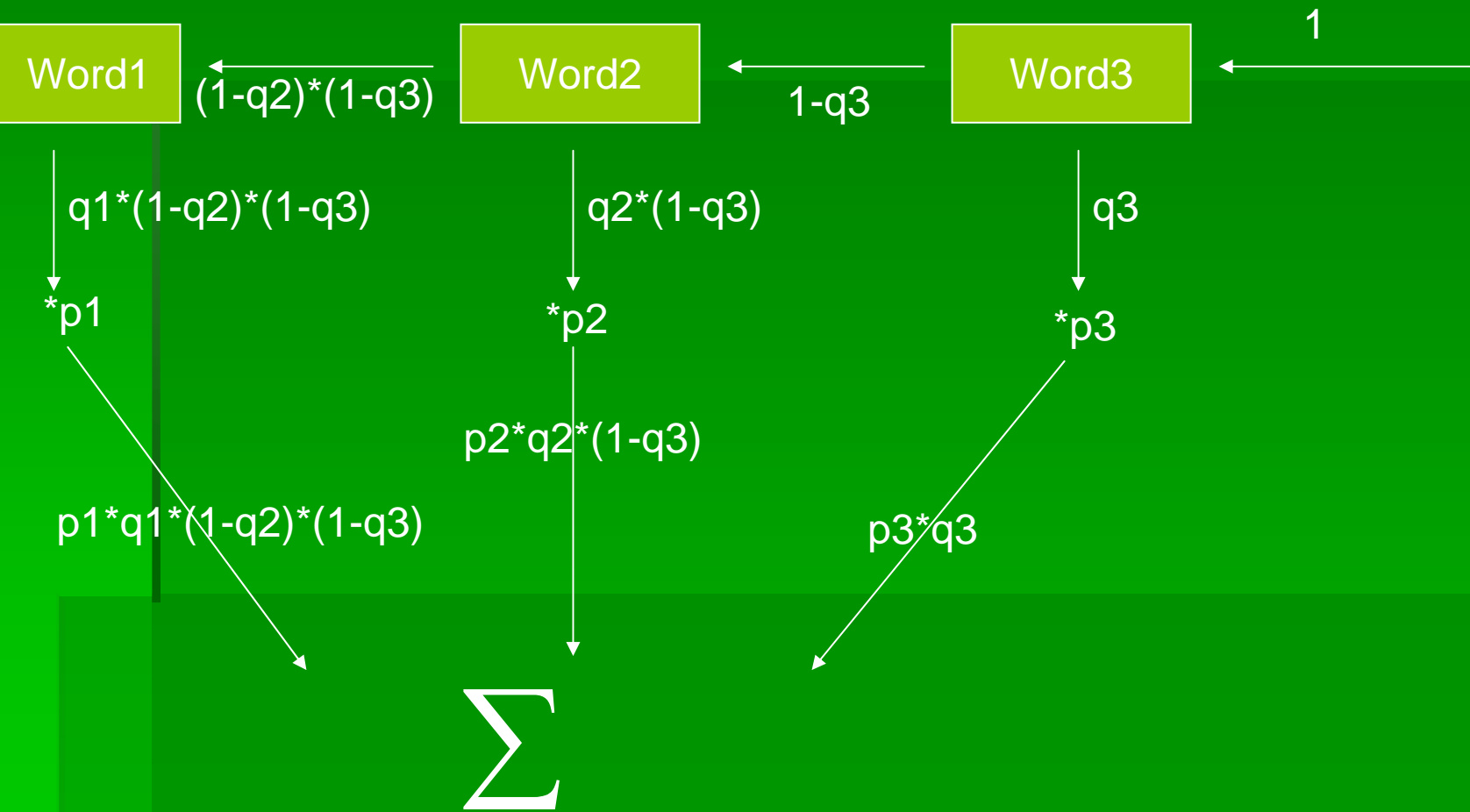
Problems

- “rat liver alkaline phosphatase”
- Basic categories (POS) did not always make sense and seemed to explain missclassifications.
- Attempts to force basic categories.

Priority Model

- Every word that appears in a name is assigned two probabilities
- p – the probability this word indicates a gene/protein name
- q – the reliability of this word as an indicator of category

Priority Model



Formula for Probability

$$n = t_{\alpha(1)} t_{\alpha(2)} \cdots t_{\alpha(k)}$$

$$p(C_1 | n) = p_{\alpha(1)} \prod_{j=2}^k (1 - q_{\alpha(j)}) + \sum_{i=2}^k q_{\alpha(i)} p_{\alpha(i)} \prod_{j=i+1}^k (1 - q_{\alpha(j)}).$$

Optimization

$$F = \sum_{h \in T_1} \log(p(C_1 | n)) + \sum_{h \in T_2} \log(p(C_2 | n))$$

Limited Memory Quasi-Newton method (L-BFGS)
Nash and Nocedal (1991)

Method	Run	P	R	F
PCFG-3	gp1	0.883	0.934	0.908
	gp2	0.882	0.937	0.909
	gp3	0.877	0.936	0.906
PCFG-8	gp1	0.939	0.966	0.952
	gp2	0.938	0.967	0.952
	gp3	0.939	0.966	0.952
LM	gp1	0.920	0.968	0.944
	gp2	0.923	0.968	0.945
	gp3	0.917	0.971	0.943
PM	gp1	0.949	0.968	0.958
	gp2	0.950	0.968	0.960
	gp3	0.950	0.967	0.958

Unknown Strings

- Variable Order Markov Model for Strings
- Used in all approaches

Markov Model

$$p(C | x_1 x_2 x_3 \dots x_n) = \frac{p(x_1 x_2 x_3 \dots x_n | C) p(C)}{p(x_1 x_2 x_3 \dots x_n)}$$

$$p(x_1 x_2 x_3 \dots x_n | C) = \prod_{k=1}^n p(x_k | x_1 x_2 x_3 \dots x_{k-1}, C)$$

Estimation

$$p(x_k | x_1 x_2 x_3 \dots x_{k-1}, C)$$

$$\{(s_i, p_i)\}_{i=1}^M$$

Estimation

$r \geq 1$ smallest $r \ni x_r x_{r+1} x_{r+2} \dots x_k$ appears in s_i for some i .

$$N' = \{i \mid x_r x_{r+1} x_{r+2} \dots x_k \text{ appears in } s_i\}$$

$$N = \{i \mid x_r x_{r+1} x_{r+2} \dots x_{k-1} \text{ appears in } s_i\}$$

$$p(x_k \mid x_1 x_2 x_3 \dots x_{k-1}, C) = \frac{\sum_{i \in N'} p_i}{\sum_{i \in N} p_i}$$

<i>GP</i>	
<u>!apoe</u>	9.55×10^{-7}
oe- <u>e</u>	2.09×10^{-3}
e- <u>epsilon</u>	4.00×10^{-2}
<i>p(apoe-epsilon GP)</i>	7.98×10^{-11}
<i>NGP</i>	
<u>!apoe</u>	8.88×10^{-8}
poe- <u>e</u>	1.21×10^{-2}
oe- <u>e</u>	6.10×10^{-2}
e- <u>epsilon</u>	6.49×10^{-3}
<i>p(apoe-epsilon NGP)</i>	4.25×10^{-13}

Extension of Priority Model

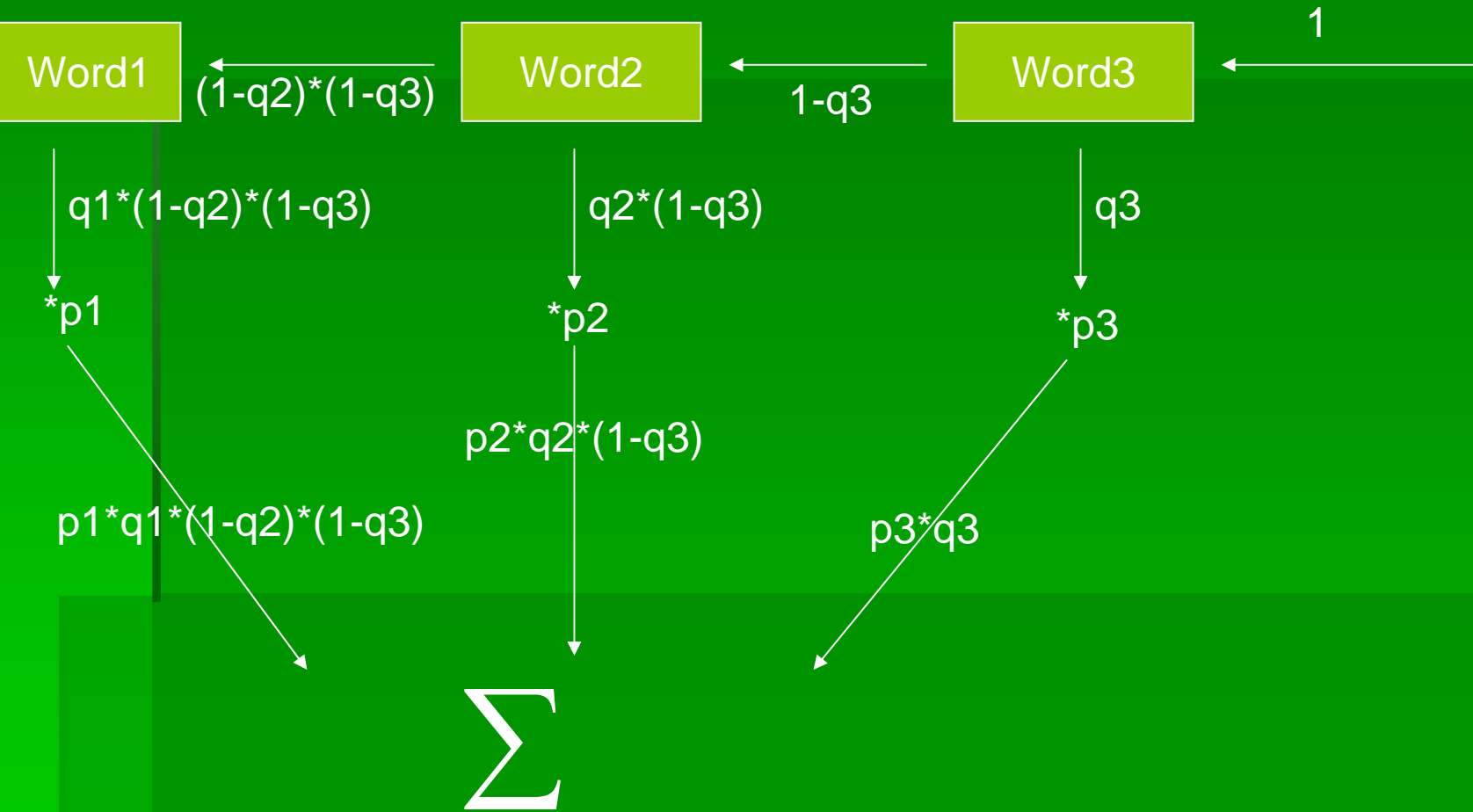
$$\left\{ (token_i, p_i) \right\}_{i=1}^M \quad \left\{ (token_i, 1 - p_i) \right\}_{i=1}^M$$
$$\left\{ (token_i, q_i) \right\}_{i=1}^M \quad \left\{ (token_i, 1 - q_i) \right\}_{i=1}^M$$

Markov Model

$$p(C | x_1 x_2 x_3 \dots x_n) = \frac{p(x_1 x_2 x_3 \dots x_n | C) p(C)}{p(x_1 x_2 x_3 \dots x_n)}$$

$$p(x_1 x_2 x_3 \dots x_n | C) = \prod_{k=1}^n p(x_k | x_1 x_2 x_3 \dots x_{k-1}, C)$$

Priority Model



Acknowledgments

- Lorrie Tanabe
- Lynne Thom
- Wayne Matten
- Don Comeau