Content Selection: "CLASSY" & Discourse

NLP Systems & Applications

LING 573

April 10, 2018

Announcements

- Start-of-quarter survey online; thanks to those who have answered already.
- For D2, the "unique_alphanumeric" should be your group number

Begin Recording!

Roadmap

- Refresher on LLR, Statistical Significance
- Content Selection
  - “CLASSY”: HMM methods
  - Discourse structure
    - Models of discourse structure
    - Structure and relations for summarization
- MEAD Demo (Maybe)

LLR & Term Significance Revisited

Log Likelihood Ratio (LLR)

$$
\begin{aligned}
LLR &= \log\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right) \\
&= \log(\text{likelihood for null model}) - \log(\text{likelihood for alternative model})
\end{aligned}
$$

- null model
  - the word *w* occurs **equally** in general as in topic
- alternative model
  - the word *w* is **more salient** for the topic than in general

Refresher on Distribution

- We can think of a document collection X as a random variable
  - Our estimate for $p(w)$ can be thought of as a sampling from X
  - (A summation of Bernoulli trials… did we see the word or not at position i? Yes or no.)

- We can thus also think of our counts for w as a ***dependent variable***
  - Where the choice of document set is the ***independent variable***

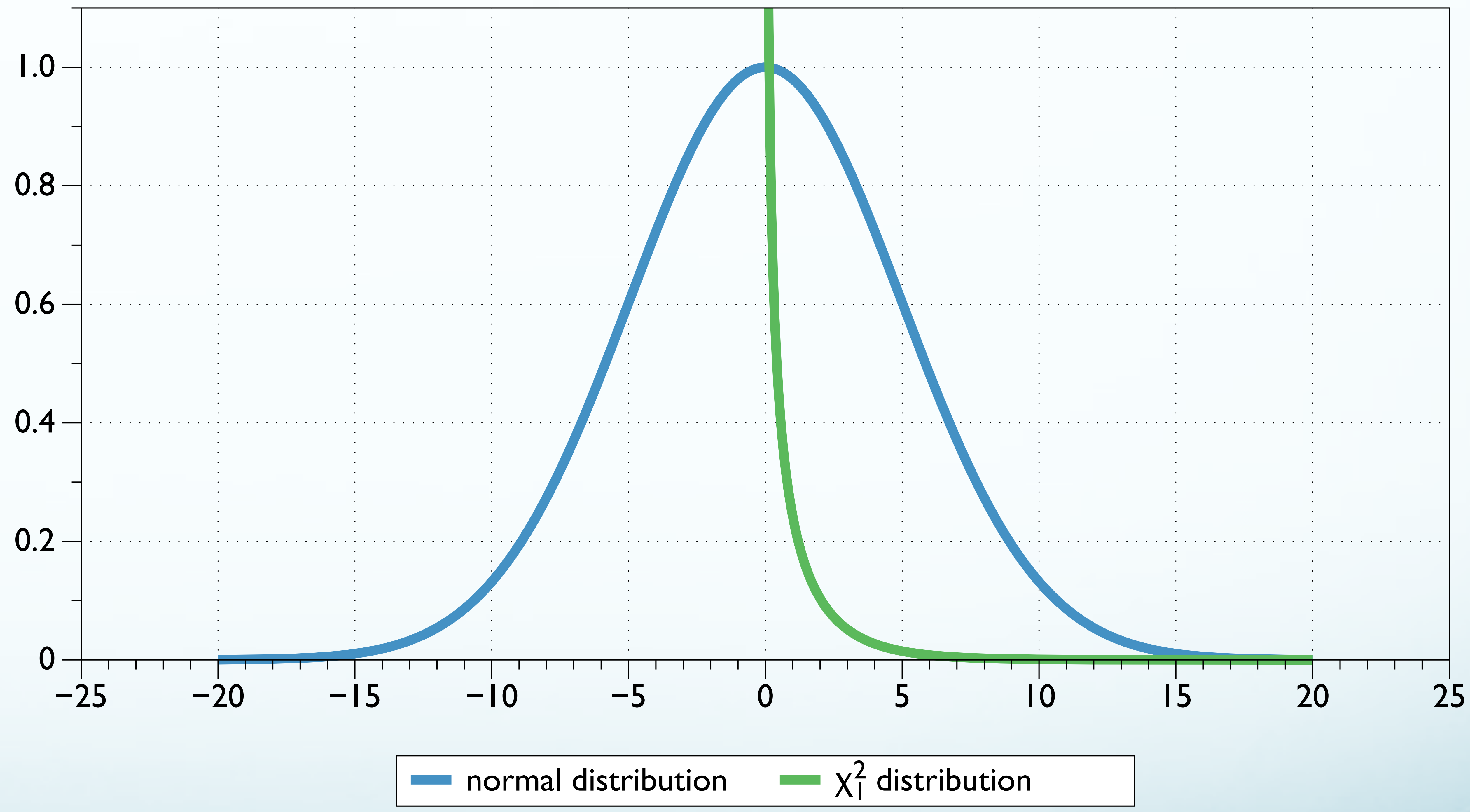
Refresher on Distribution

- Two document sets → two independent variables, D_1 and D_2
- Hypotheses:
  - **H_0** — w's distribution behaves **the same** between choices of independent variable
  - **H_1** — w's distribution behaves **differently** between choices of independent variable
- We can reason about which hypothesis is more likely based upon:
  - How often do we **expect** to observe w in a document set?
  - How often do we **observe** w in a document set?

Intuitive Understanding of χ^2

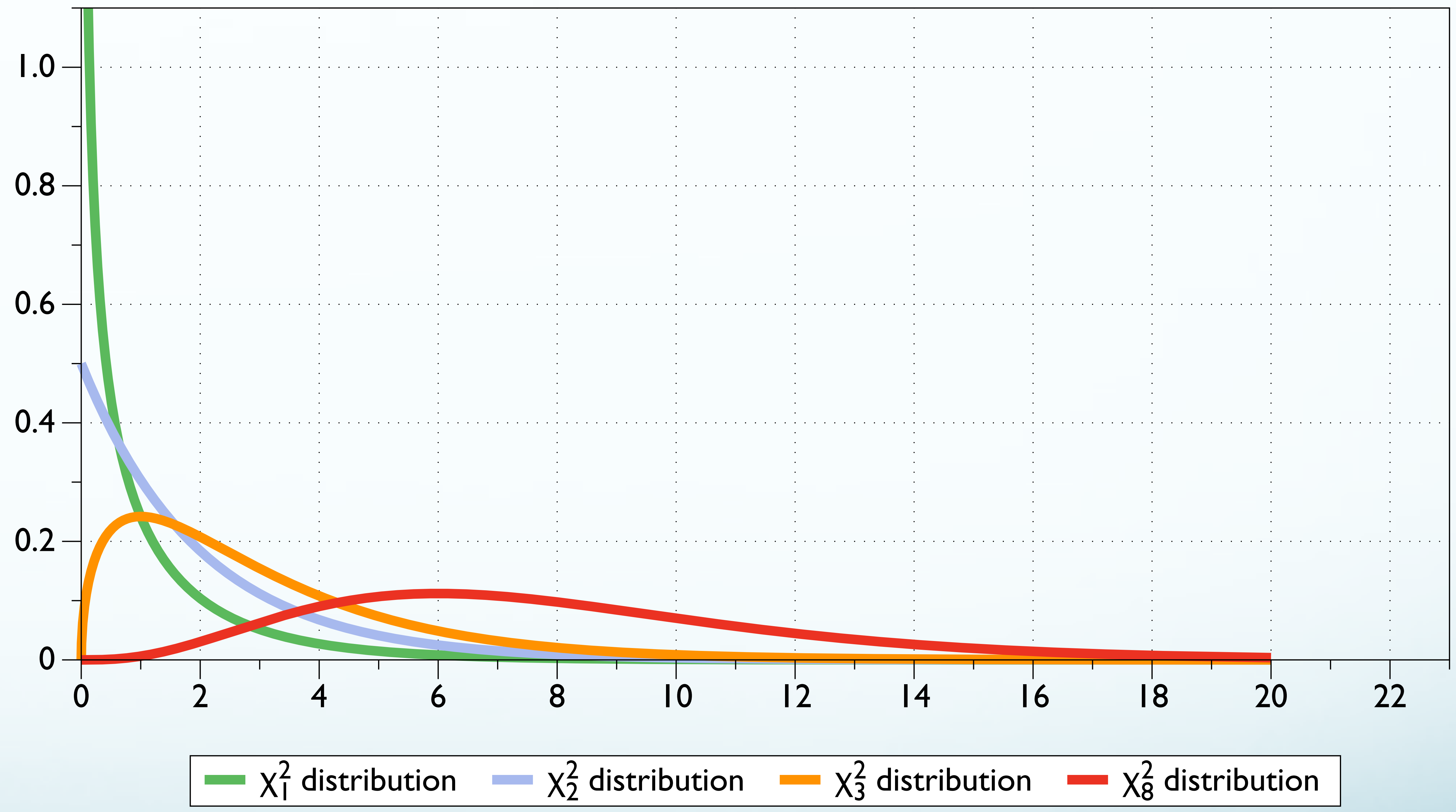
- A χ^2 distribution shows how likely it

  is to find a proportion of samples some distance from the mean.
  - (Its values are only non-negative)
- One degree of freedom
  - 2 independent variables
  - the odds of sampling around the mean are high.



Intuitive Understanding of χ^2

- The more degrees of freedom (different independent variables)
  - The more likely that something will deviate from the mean just by random fluctuation in the different variables



Intuitive Understanding of χ^2

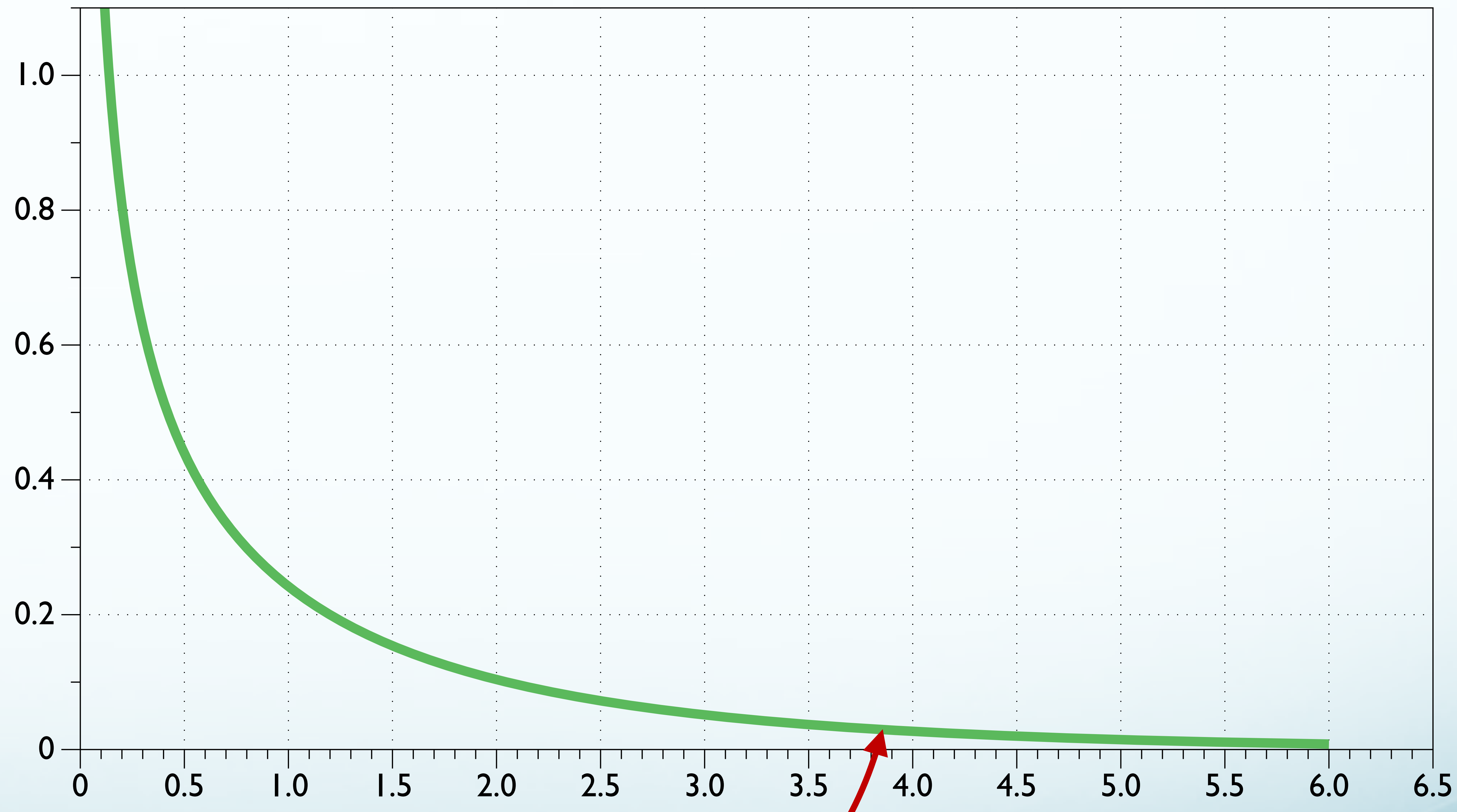
- This makes sense:
  - If the random variables are attributes of a dog:
    - Weight, Height, Leg Length, Torso Length, Tail Thickness, Fur Length, Ear Height
    - …what is the likelihood that any given sample of "dog" will resemble the mean of all of these characteristics?

Average

(via Kumar & Singh, 2014)

Intuitive Understanding of χ^2

- Knowing the degrees of freedom for the problem, we can see the probability of a sampling that far off from the expected mean is.
- For 1 df, and χ^2 value 3.84
  - We have a 5% chance of getting a result this far off the norm
  - (a *p*-value of 0.05)



Log Likelihood Ratio (LLR)

- w **outside topic**:
  - $k_b =$ count of w outside topic
  - $n_b =$ total words outside topic

$$p_b = \frac{k_b}{n_b}$$

- w **in topic**:
  - $k_t =$ count of w in topic
  - $n_t =$ total words in topic

$$p_t = \frac{k_t}{n_t}$$

- w **overall**:
  - $k_o =$ count of w overall (k_{topic} + $k_{background}$)
  - $n_o =$ total words overall (n_{topic} + $n_{background}$)

$$p_t = \frac{k_o}{n_o}$$

Log Likelihood Ratio (LLR)

- Likelihood of a model, from frame of word w
  - Product of prob p of seeing word w by times seen k $\quad p^k$
  - Product of seeing all other words in corpus of size n $\quad (1-p)^{n-k}$
  - …times the number of ways this sequence could happen: $\binom{n}{k}$

$$\text{likelihood for alternative model} = \binom{n_t}{k_t}(p_t)^{k_t}\cdot(1-p_t)^{(n_t-k_t)}\cdot\binom{n_b}{k_b}(p_b)^{k_b}\cdot(1-p_b)^{(n_b-k_b)}$$

$$\text{likelihood for null model} = \binom{n_t}{k_t}(p_o)^{k_t}\cdot(1-p_o)^{(n_t-k_t)}\cdot\binom{n_b}{k_b}(p_o)^{k_b}\cdot(1-p_o)^{(n_b-k_b)}$$

Log Likelihood Ratio (LLR)

- Example: we see the word "train"
  - 10 times in a topic of 100 words
  - 2 times outside the topic, with 200 words

$$p_t = \frac{10}{100} = 0.10 \qquad p_b = \frac{2}{200} = 0.01 \qquad p_o = \frac{12}{300} = 0.04$$

Log Likelihood Ratio (LLR) — H1

in-topic likelihood $\ln\left[\binom{100}{10}\cdot 0.10^{10}\cdot(1-0.10)^{(100-10)}\right] = \ln\binom{100}{10} + 10\times\ln(0.10) + 90\times\ln(0.9)$

$= -2.0259$

out-topic likelihood $\ln\left[\binom{200}{2}\cdot 0.01^{2}\cdot(1-0.01)^{(200-2)}\right] = \ln\binom{200}{2} + 2\cdot\ln(0.01) + 198\cdot\ln(0.99)$

$= -1.302$

Log Likelihood Ratio (LLR) — H0

in-topic likelihood

$$\ln\left[\binom{100}{10}\cdot 0.04^{10}\cdot(1-0.04)^{(100-10)}\right]=\ln\binom{100}{10}+10\times\ln(0.04)+90\times\ln(0.96)$$

$$=-5.380$$

out-topic likelihood

$$\ln\left[\binom{200}{2}\cdot 0.04^{2}\cdot(1-0.04)^{(200-2)}\right]=\ln\binom{200}{2}+2\cdot\ln(0.04)+198\cdot\ln(0.96)$$

$$=-4.622$$

Log Likelihood Ratio (LLR)

$$\begin{aligned} LLR &= \text{LL for null model - LL for alternative model} \\ &= (-5.380 - 4.622) - (-2.026 - 1.302) \\ &= -6.675 \end{aligned}$$

- Using the base of e from the $\ln$: $e^{-6.675} = 0.00126$
- Meaning the likelihood of the null hypothesis is $(0.00126)\times 100 = 0.126\%$ as likely as the alternative hypothesis

Significance testing LLR

- Significance tests, such as Chi-squared are typically used with LLR as well
- $-2 \times LLR$ is the test statistic used, called D, –2LL, or –2logλ
- Given that we had a distribution that could be represented by the contingency table:

	Inside Topic	Outside Topic
word w	p_t	p_b
other word	1-p_t	1-p_b

- We can consider this to have **one** degree of freedom, and can use χ^2 table:

Confidence Value	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005	0.001
Threshold	0	0	0	0	0.02	2.71	3.84	5.02	6.63	7.88	10.83

Significance testing LLR

- So for our example, $-2 \times -6.675 =$ **13.35**

Confidence Value	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005	0.001
Threshold	0	0	0	0	0.02	2.71	3.84	5.02	6.63	7.88	10.83

- So our statistical significance is well above $p < 0.001$

“CLASSY”

Conroy et al (2001, 2004, …)

CLASSY

- "Clustering, Linguistics, and Statistics for Summarization Yield"
  - Conroy et al. 2000—2011 (2001, 2004, 2006)
- Highlights:
  - High performing system
    - Often rank 1 in DUC/TAC, commonly used comparison
  - Topic signature-type system (LLR)
  - Two approaches to content selection:
    - Matrix Decomposition
    - HMM
  - Redundancy handling

CLASSY

- **Key assumption**:
  - Informitiveness of sentence is largely determined by number of salient words
  - Best sentences for selection will be maximally informitive
- **Two approaches**:
  - Matrix decomposition
  - HMM

CLASSY — Matrix Decomposition

- Frames content selection as a linear algebra problem
- Pivoted QR Decomposition
  - The columns of the **R** matrix end up representing the sentences ranked in order of importance

$$\underset{\text{A}}{\begin{pmatrix} \vdots & \vdots & \vdots \\ a_1 & a_2 & a_3 \\ \vdots & \vdots & \vdots \end{pmatrix}} = \underset{\text{Q}}{\begin{pmatrix} \vdots & \vdots & \vdots \\ e_1 & e_2 & e_3 \\ \vdots & \vdots & \vdots \end{pmatrix}} \underset{\text{R}}{\begin{pmatrix} e_1^T \cdot a_1 & e_1^T \cdot a_2 & e_1^T \cdot a_3 \\ & e_2^T \cdot a_2 & e_2^T \cdot a_3 \\ & & e_3^T \cdot a_3 \end{pmatrix}}$$

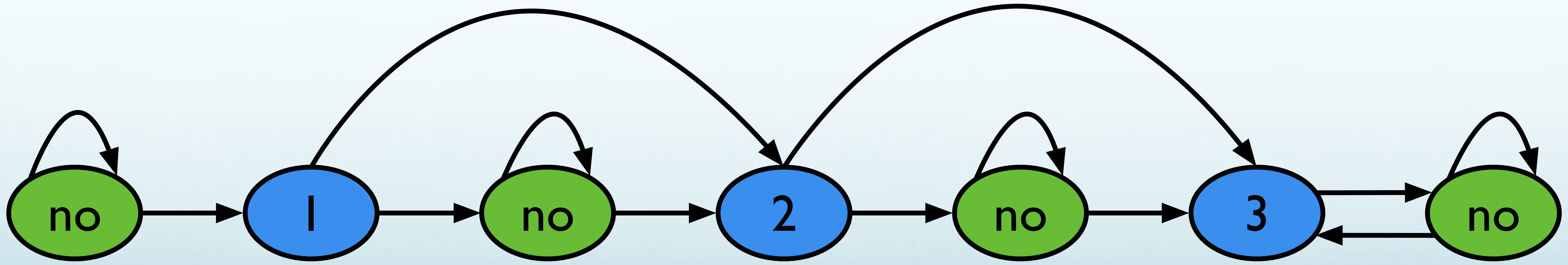
Original Matrix | Orthogonal Unit Vectors | Upper Diagonal Matrix

CLASSY — Matrix Decomposition

- Redundancy minimizing selection
- Create [term×sentence] matrix — If term is in sentence, weight is nonzero
- Loop:
  - Select highest scoring sentence
    - Based on Euclidean norm (magnitude of sentence vector)
  - Subtract those components (representing terms) from remaining sentences
  - Until enough sentences
- **Effect**: selects highly ranked but different sentences
  - Relatively insensitive to weighting schemes

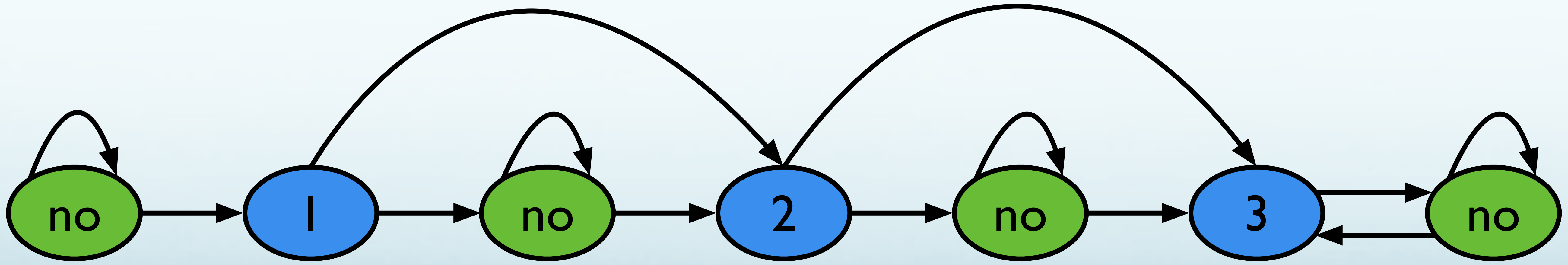
HMM Sentence Selection

- Intuition:
  - Summarization can be thought of as a sequence labeling task
  - Between labels that correspond to different "reasons" for **inclusion** or **exclusion**
  - Additionally captures positional information
    - How likely is a highly "contentful" sentence to be followed by one equally contentful?



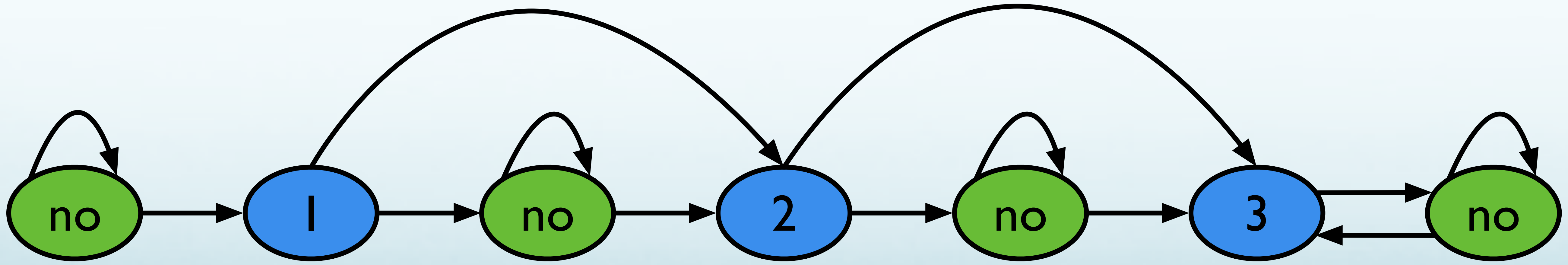
HMM Sentence Selection

- Features attempted for HMM:
  - Position of sentence in document
  - Position of sentence in paragraph
  - Number of terms in sentence
  - **LLR**



HMM Sentence Selection

- CLASSY strategy: Use **LLR** to represent sentences in HMM
  - Two **classes** of states (13 states total, "empirically determined")
    - **Include this sentence!**
    - **Don't include this sentence.**
  - Trained on human summaries of docsets
  - System must go through three "**lead**" states, then can loop.



Combining Approaches

- Both HMM and Matrix method select sentences
- Can combine to further improve
- Approach:
  - Use HMM method to compute sentence scores
    - (e.g. rather than just weight based)
      - Incorporates context information, prior states
  - Loop:
    - Select highest scoring sentence
    - Update matrix scores
      - Exclude those with too low matrix scores
    - Until enough sentences are found

Other Linguistic Processing

- Sentence manipulation (before selection):
  - Remove uninteresting phrases based on POS tagging
    - Gerund clauses, appos, attrib, lead adverbs
- Coreference handling (Serif system)
  - Created coref chains initially
  - Replace all mentions with longest mention (# capital letters)
  - Used only for sentence selection

Outcomes

- HMM, Matrix: both effective, better combined
- Linguistic pre-processing improves
  - Best ROUGE-1, ROUGE-2 in DUC
- Coref handling improves
  - Best ROUGE-3, ROUGE-4; 2nd ROUGE-2

Discourse Structure for Content Selection

Louis et. al (2010)

Discourse Relations

- Discourse relations:
  - Possible meaning relations between utterances in discourse
  - Examples:
    - **Result**: Infer state of S_0 causes state in S_1
      - The Tin Woodman was caught in the rain. His joints rusted.
    - **Explanation**: Infer state in S_1 caused state in S_0
      - John hid Bill's car keys. He was drunk.
    - **Elaboration**: Infer same prop. from S_0 and S_1.
      - Dorothy was from Kansas. She lived in the great Kansas prairie.
- Pair of locally coherent clauses: **discourse segment**

Discourse Structure for Content Selection

- Key Intuitions:
  - Different discourse relations have different relevance for inclusion in summary
    - e.g. **elaboration** likely less helpful than **result** or **explanation**
  - Structure — Some information more "core"
    - nucleus vs. satellite, promotion, centrality

Rhetorical Structure Theory

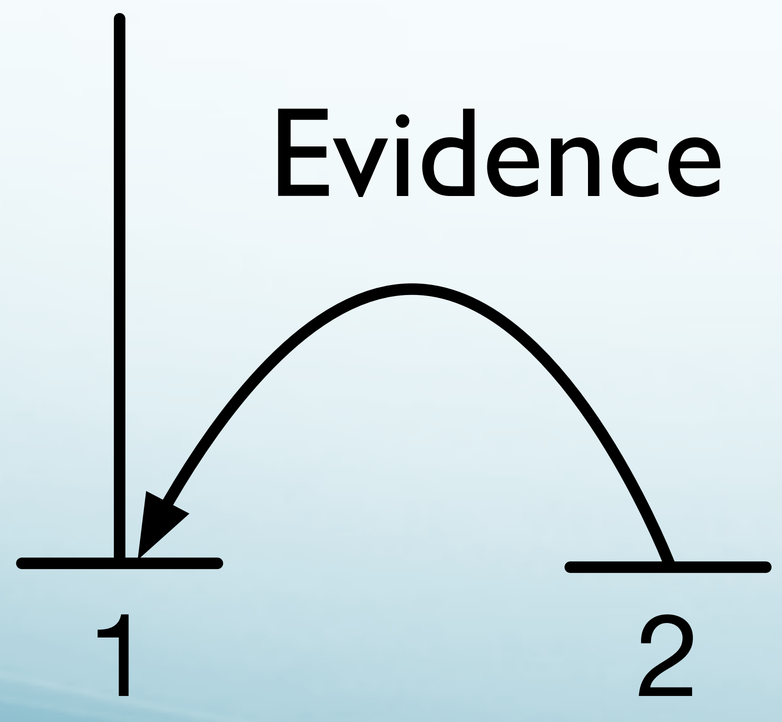
- Mann & Thompson (1988)
- Goal: Identify hierarchical structure of text
  - Cover wide range of text types
    - Language contrasts
  - Relational propositions (intentions)
- Derives from functional relations b/t clauses

Components of RST

- Relations hold between two text spans, **nucleus** and **satellite**
  - Nucleus core element, satellite peripheral
  - Constraints on each, between
  - Units: Elementary discourse units (EDUs), e.g. clauses

RST Relations

- **Evidence**
  - *The program really works.* (N)
  - *I entered all my info and it matched my results.* (S)



Relation Name:	Evidence
Constraints on N:	R might not believe N to a degree satisfactory to W
Constraints on S:	R believes S or will find it credible
Constraints on N+S:	R's comprehending S increases R's belief of N
Effects:	R's belief of N is increased

RST Relations

- Core of RST
  - RST analysis requires building tree of relations
  - Relations include
    - Circumstance, Solutionhood, Elaboration, Background, Enablement, Motivation, Evidence, etc.
- Captured in:
  - RST treebank: corpus of WSJ articles with analysis (`/corpora/LDC/LDC02T07` on Patas)
  - RST parsers: Marcu 1996, Feng and Hirst 2014

GraphBank

- Alternative discourse structure model
  - Wolf & Gibson, 2005
- Key difference:
  - Analysis of text need not be tree-structure, like RST
  - Can be arbitrary graph, allowing crossing dependencies
- Similar relations among spans (clauses)
  - Slightly different inventory

Penn Discourse Treebank

- PDTB (Prasad et al, 2008)
  - "Theory-neutral" discourse model
  - No stipulation of overall structure, identifies local relations only
- Two types of annotation:
  - Explicit — lexical markers such as "because," "but," "while,"
  - Implicit — No explicit lexical markers, more like RST examples
- Senses/Relations:
  - Comparison, Contingency, Expansion, Temporal…

Other Thoughts for Discourse

- Also useful for **information ordering**:
  - e.g. Make sure that nucleus is introduced before satellites
- **Realization**:
  - That sequential sentences are coherent, in additional to cohesive
- Compare these, contrast, with lexical info alone
  - Louis et al, 2010

More About Discourse

Framework

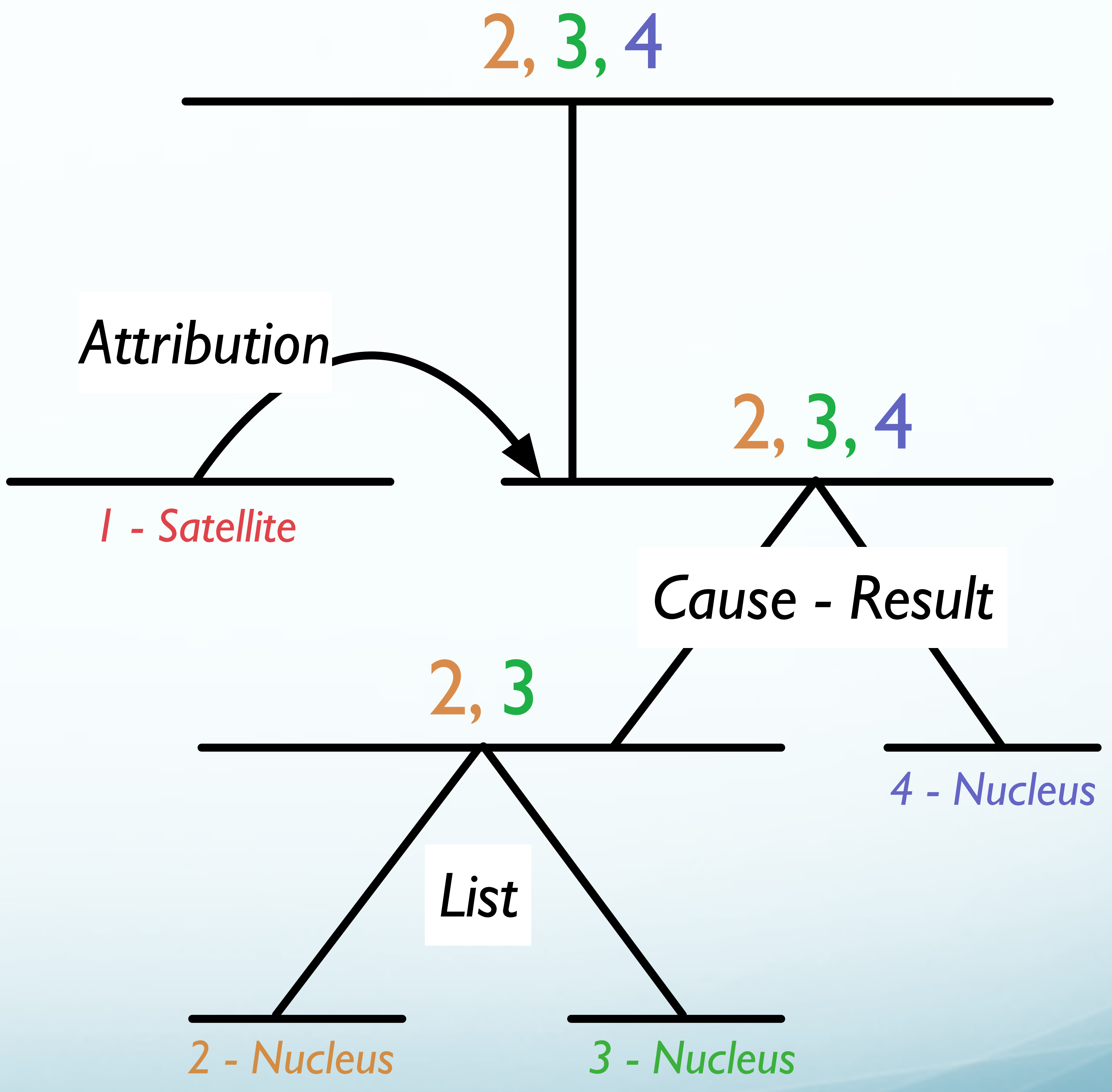
- Association with extractive summary sentences
  - Statistical analysis
    - χ^2 (categorical) t-test (continuous)
- Classification:
  - Logistic regression
    - Different ensembles of features
  - Classification F-measure
  - ROUGE over summary sentences

RST Parsing

- Learn and apply classifiers for
  - Segmentation and parsing of discourse
- Assign coherence relations between spans
- Create a representation over whole text → parse
- Discourse structure
  - RST trees
    - Fine-grained, hierarchical structure
      - Clause-based units

Discourse Structure Example

- 1. [Mr. Watkins said]
- 2. [volume on Interprovincial's system is down about 2% since January]
- 3. [and is expected to fall further,]
- 4. [making expansion unnecessary until perhaps the mid-1990s.]

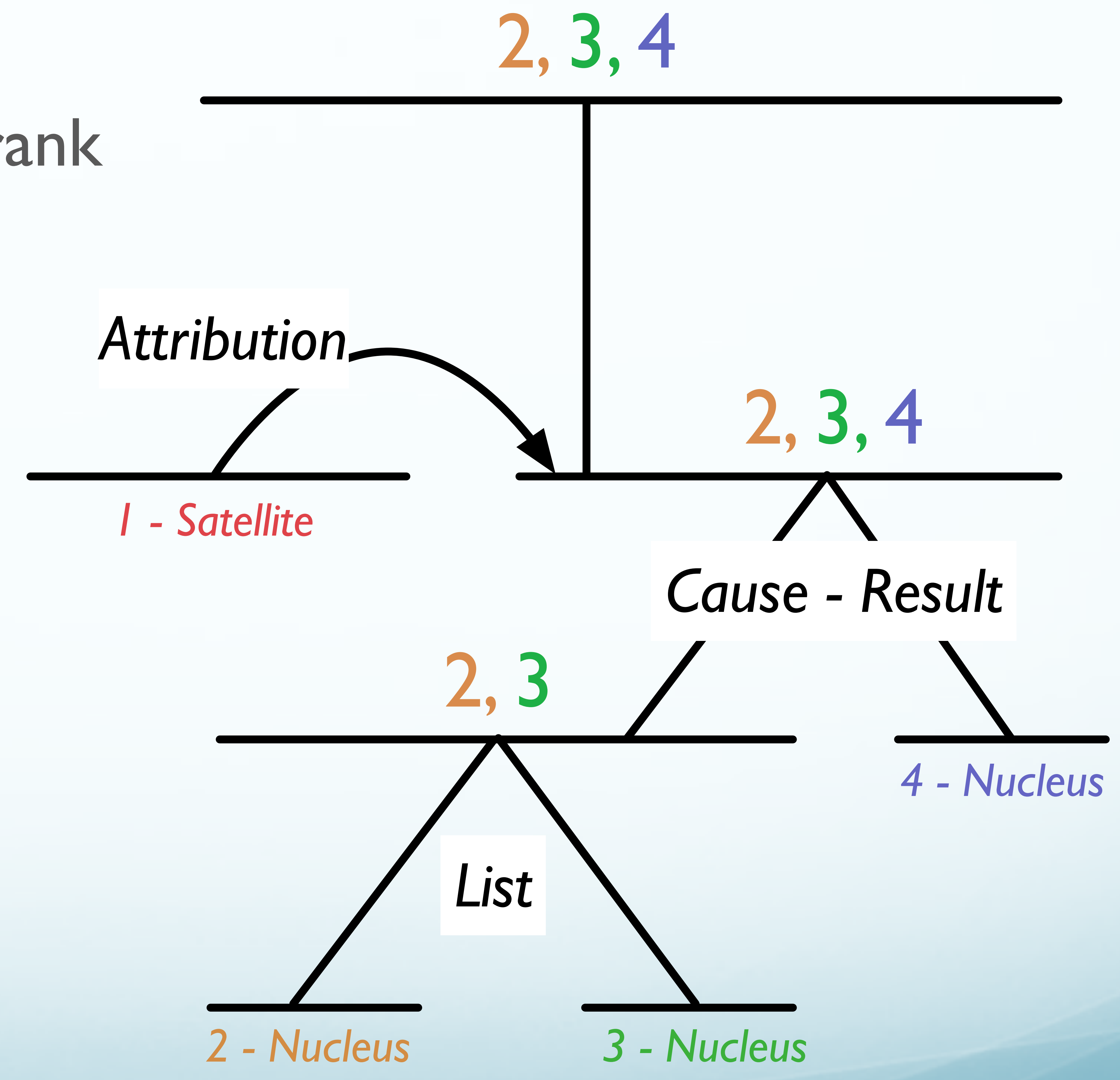


Discourse Structure Features

- **Satellite penalty**
  - For each EDU — number of satellite nodes between EDU and root
    - 1 satellite in tree: one step to root: penalty = 1
- **Promotion set**:
  - Nuclear units at some level of tree
    - At leaves, EDUs are themselves nuclear

Discourse Structure Features

- **Depth score**:
  - Distance from lowest tree level to EDU's highest rank
    - **2,3,4**: score=4
    - **1**: score=3
- **Promotion score**:
  - # of levels span is promoted
    - **1**: score = 0
    - **4**: score = 2
    - **2,3**: score = 3



Converting to Sentence Level

- Each feature has:
  - Raw score
  - Normalized score: $\dfrac{\text{raw score}}{\text{sentence length}}$
- Sentence score for a feature:
  - Maximum value over all EDUs in sentence

"Semantic" Features

- Capture specific relations on spans
- Binary features over tuple of:
  - Implicit vs. Explicit
  - Name of relation between units
  - If a relation exists between sentences:
    - Whether sentence is Arg1 or Arg2
- Also:
  - Number of relations
  - Distance between arguments within sentence

Example 1

- *In addition, its machines are easier to operate, so customers require less assistance from software.*

- Is there an explicit discourse marker?
  - Yes, "**so**"

- Discourse relation?
  - ***Contingency***

Example II

- *(1) Wednesday's dominant issue was Yasuda & Marine Insurance, which continued to surge on rumors of speculative buying. (2) It ended the day up 80 yen to 1880 yen.*

- Is there an explicit discourse marker?
  - **No**

- Is there a relation?
  - **Yes, Implicit**.

- What relation?
  - **Expansion**. (More specifically, restatement).

Non-Discourse Features

- Typical Features
  - Sentence length
  - Sentence position
  - Probabilities of words in sentence: mean, sum, product
  - # of signature words (LLR)

Significant Features: Summary Sentences

- Structure:
  - depth score
  - promotion score
- Semantic:
  - Arg1 of Explicit Expansion
  - Implicit Contingency
  - Implicit Expansion
  - Distance to Arg
- Non-discourse:
  - length
  - 1st in paragraph
  - offset from end of paragraph
  - # signature terms
  - mean
  - sum word probabilities

Significant Features: Non-Summary Sentences

- Structure:
  - satellite penalty
- Semantic:
  - Explicit expansion
  - Explicit contingency
  - Arg2 of implicit temporal
  - Arg2 of implicit contingency
  - # of shared relations
- Non-discourse:
  - offset from paragraph start
  - offset from article start
  - sentence probability

Observations

- Non-discourse features good cues to summary
- Structural features match intuition
- Semantic features
  - Relatively few useful features for selecting summaries
  - Most features associated with non-summary… but most sentences are non-summary

Evaluation

- Structural is best, both alone and in combination
- Best overall combines all types
- Both F_1 and ROUGE

Features used	Acc	P	R	F
structural	78.11	63.38	22.77	33.50
semantic	75.53	44.31	5.04	9.05
non-discourse (ND)	77.25	67.48	11.02	18.95
ND + semantic	77.38	59.38	20.62	30.61
ND + structural	78.51	63.49	26.05	36.94
semantic + structural	77.94	58.39	30.47	40.04
structural + semantic + ND	78.93	61.85	34.42	44.23