CS447: Natural Language Processing
http://courses.engr.illinois.edu/cs447

Lecture 24: A very brief introduction to discourse

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

Projects and Literature Reviews

First report due Nov 26
(PDF written in LaTeX; no length restrictions; submission through Compass)

Purpose of this first report:
Check-in to make sure that you're on track
(or, if not, that we can spot problems)

Rubrics for the *final* reports (due on Reading Day):
https://courses.engr.illinois.edu/CS447/LiteratureReviewRubric.pdf
https://courses.engr.illinois.edu/CS447/FinalProjectRubric.pdf

Projects and Literature Reviews

Guidelines for first Project Report:
- What is your project about?
- What are the relevant papers you are building on?
- What data are you using?
- What evaluation metric will you be using?
- What models will you implement/evaluate?
- What is your to-do list?

Guidelines for first Literature Review Report:
- What is your literature review about?
- (What task or what kind of models?
- Do you have any specific questions or focus?)
- What are the papers you will review?
- (If you already have it, give a brief summary of each of them)
- What's your to-do list?

Outlook

Lectures 25—27: Neural approaches to NLP

Lecture 28 (Wed, Dec 12): Final exam
(in-class, closed book, only materials after midterm)

Fixing my bug from the last lecture…

Finding the best translation

How can we find the *best* translation efficiently?
There is an exponential number of possible translations.

We will use a *heuristic* search algorithm
We cannot guarantee to find the best (= highest-scoring) translation, but we're likely to get close.

We will use a *"stack-based"* decoder
(If you've taken Intro to AI: this is A* ("A-star") search)
We will score partial translations based on how good we expect the corresponding completed translation to be.
Or, rather: we will score partial translations on how **bad** we expect the corresponding complete translation to be.
That is, our scores will be **costs (high=bad, low=good)**

Scoring partial translations

Assign expected costs to *partial* translations (E, F):

$$expected\_cost(E,F) = current\_cost(E,F) + future\_cost(E,F)$$

The current cost is based on the score of the partial translation (E, F)

e.g. $current\_cost(E,F) = \log P(E)P(F \mid E)$

The (estimated) future cost is a **lower** bound on the actual cost of completing the partial translation (E, F):

$true\_cost(E,F)$ $(= current\_cost(E,F) + actual\_future\_cost(E,F))$

$\geq expected\_cost(E,F)$ $(= current\_cost(E,F) + est\_future\_cost(E,F))$

because $actual\_future\_cost(E,F) \geq est\_future\_cost(E,F)$

(The estimated future cost ignores the distortion cost)

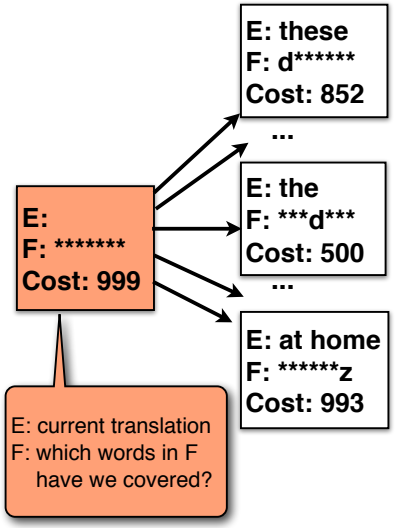
Stack-based decoding

Maintain a **priority queue** (='stack') of **partial translations** (hypotheses) with their **expected costs**.
Each element on the stack is **open** (we haven't yet pursued this hypothesis) or **closed** (we have already pursued this hypothesis)

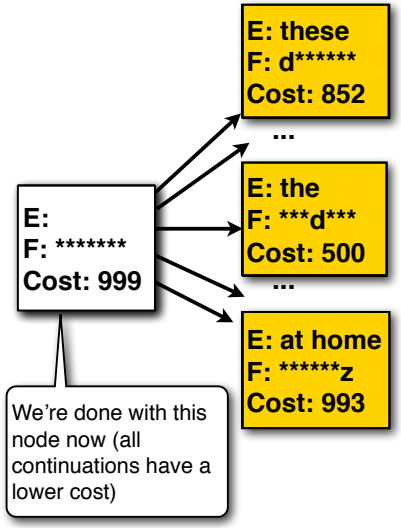
At each step:
- **Expand** the best open hypothesis (the open translation with the lowest expected cost) in all possible ways.
- These new translations become new **open elements** on the stack.
- **Close** the best open hypothesis.

Additional Pruning (*n*-best / beam search):
Only keep the *n* best open hypotheses around

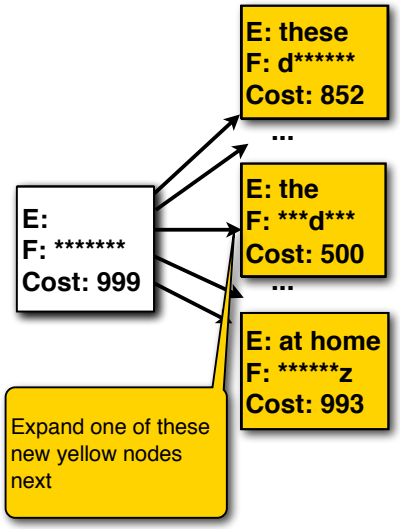
Stack-based decoding



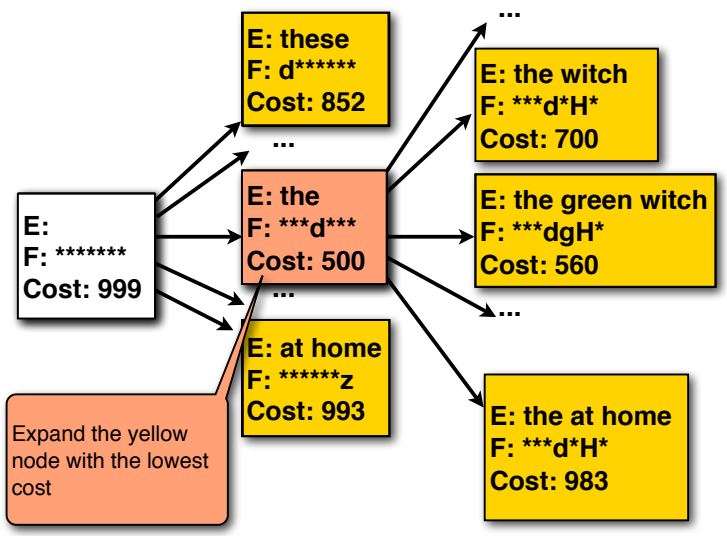
Stack-based decoding



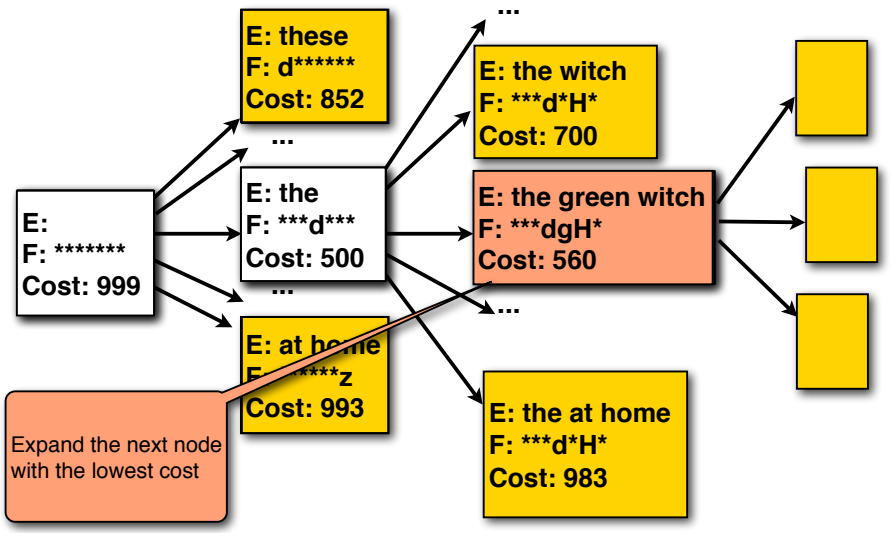
Stack-based decoding



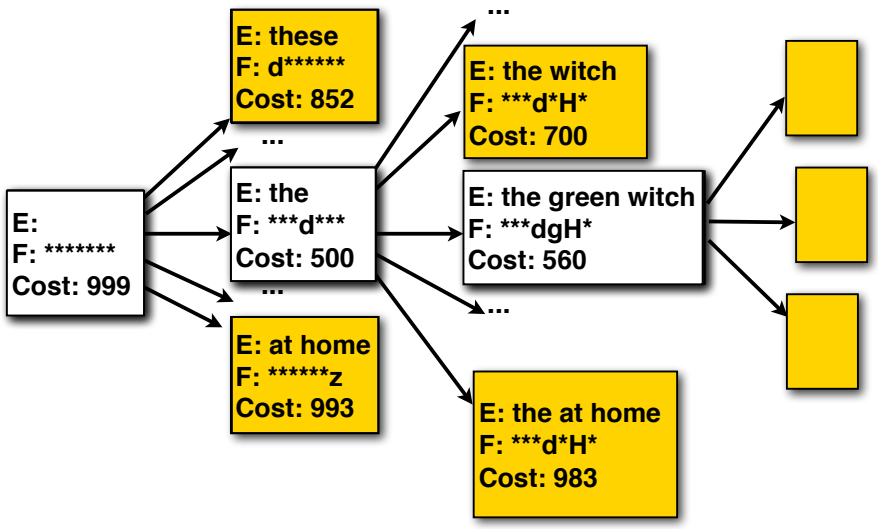
Stack-based decoding



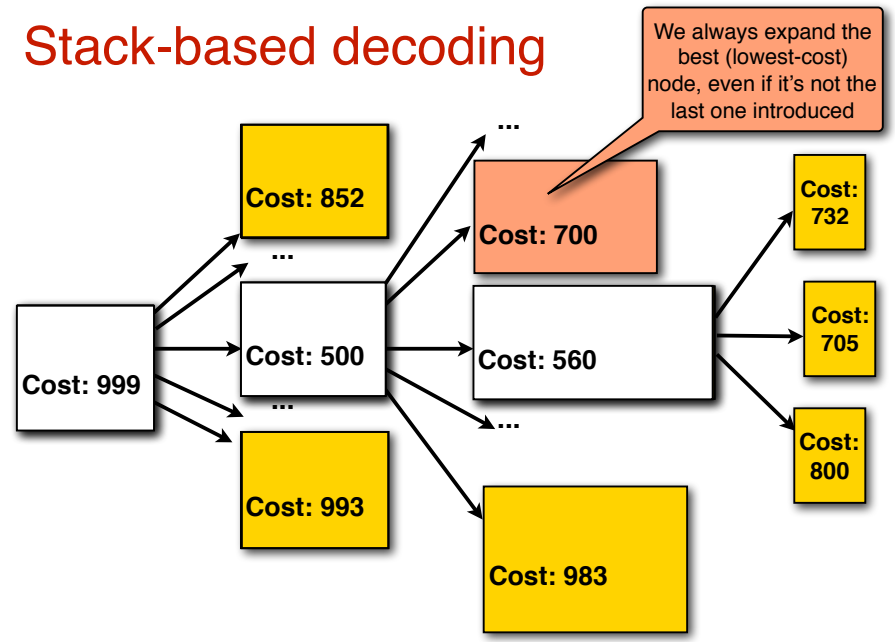
Stack-based decoding



Stack-based decoding



Stack-based decoding



Discourse

What is discourse?

On Monday, John went to Einstein's. He wanted to buy lunch. But the cafe was closed. That made him angry, so the next day he went to Green Street instead.

'Discourse':
any linguistic unit that consists of **multiple sentences**

Speakers describe "some situation or state of the real or some hypothetical world" (Webber, 1983)

Speakers attempt to get the **listener** to construct a similar **model of the situation**.

Why study discourse?

For natural language understanding:
Most information is not contained in a single sentence.
The system has to aggregate information across paragraphs or entire documents.

For natural language generation:
When systems generate text, that text needs to be easy to understand — it has to be coherent.
What makes text coherent?

How can we understand discourse?

On Monday, John went to Einstein's. He wanted to buy lunch. But the cafe was closed. That made him angry, so the next day he went to Green Street instead.

Understanding discourse requires (among other things):

1) doing coreference resolution:
'the cafe' and '*Einstein's'* refer to the same entity
He and *John* refer to the same person. *That* refers to *'the cafe was closed'*.

2) identifying discourse ('coherence') relations:
'He wanted to buy lunch' is the *reason* for 'John went to Bevande.'

Discourse models

An explicit representation of:

- the **events and entities** that a discourse talks about
- the **relations** between them (and to the real world).

This representation is often written in some form of logic.

What does this logic need to capture?

Discourse models should capture...

Physical entities: John, Einstein's, lunch

Events: On Monday, John went to Einstein's
involve entities, take place at a point in time

States: It was closed.
involve entities and hold for a period of time

Temporal relations: afterwards
between events and states

Rhetorical ('discourse') relations: ... so ... instead
between events and states

Rhetorical (Discourse) relations

Rhetorical relations

Discourse 1:
John hid Bill's car keys. He was drunk.

Discourse 2:
John hid Bill's car keys. He likes spinach.

Discourse 1 is more coherent than Discourse 2 because "He(=Bill) was drunk" provides an **explanation** for "John hid Bill's car keys"
What **kind of relations** between two consecutive utterances (=sentences, clauses, paragraphs,…) make a discourse coherent?

Rhetorical Structure Theory; also lots of recent work on **discourse parsing (Penn Discourse Treebank)**

Example: The *Result* relation

The reader can infer that the **state/event described in S0 causes** (or: could cause) **the state/event asserted in S1**:

S0: The Tin Woodman was caught in the rain.
S1: His joints rusted.

This can be rephrased as:
"S0. As a result, S1"

Example: The *Explanation* relation

The reader can infer that **the state/event in S1 provides an explanation** (reason) **for the state/event in S0**:

S0: John hid Bill's car keys.
S1: He was drunk.

This can be rephrased as:
"S0 because S1"

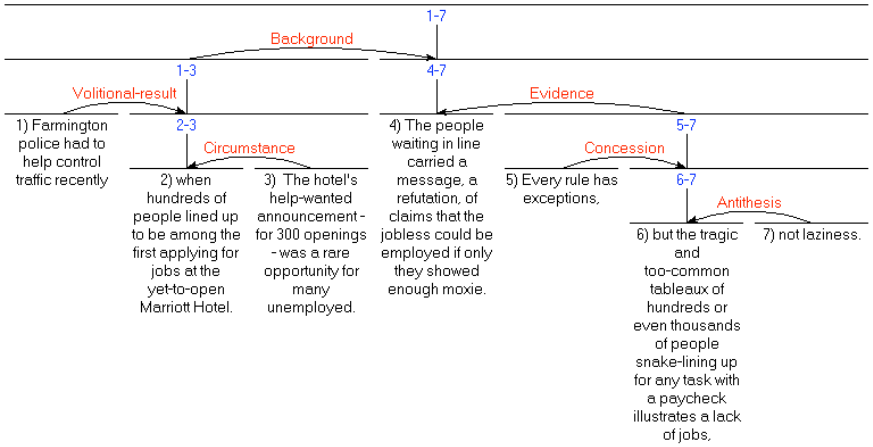
Rhetorical Structure Theory (RST)

RST (Mann & Thompson, 1987) describes **rhetorical relations** between utterances:
Evidence, Elaboration, Attribution, Contrast, List,…
Different variants of RST assume different sets of relations.

Most relations hold between a **nucleus** (N) and a **satellite** (S). Some relations (e.g. *List*) have **multiple nuclei** (and no satellite).

Every relation imposes certain **constraints** on its arguments (N,S), that describe the goals and beliefs of the **reader** R and **writer** W, and the effect of the utterance on the reader.

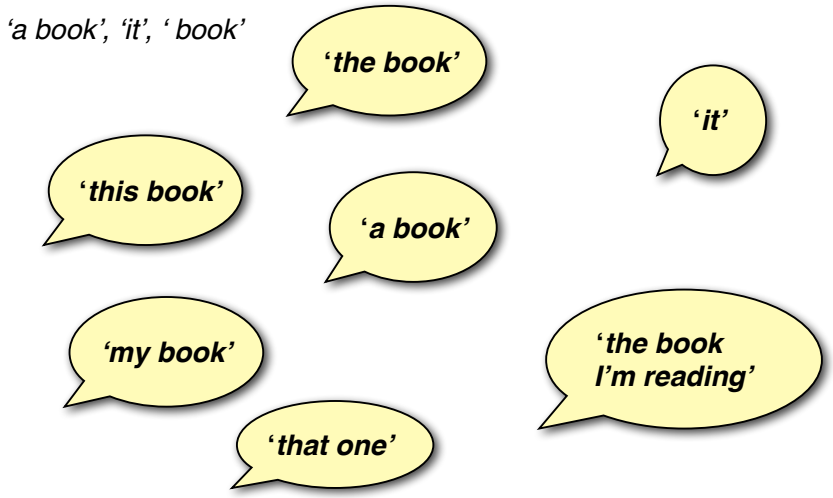
Discourse structure is hierarchical



RST website: http://www.sfu.ca/rst/

Referring expressions and coreference resolution

How do we refer to entities?



Some terminology

Referring expressions ('*this book*', 'it') refer to some entity (e.g. a book), which is called the **referent.**

Co-reference: two referring expressions that refer to the same entity **co-refer** (are co-referent).
I saw a movie last night. I think you should see it too!

The referent is **evoked** in its first mention, and **accessed** in any subsequent mention.

Indefinite NPs

- **no determiner:**
  *I like **walnuts**.*
- **the indefinite determiner:**
  *She sent her **a** beautiful goose*
- **numerals:**
  *I saw **three** geese.*
- **indefinite quantifiers:**
  *I ate **some** walnuts.*
- **(indefinite) *this*:**
  *I saw **this** beautiful Ford Falcon today*

Indefinites usually **introduce a new discourse entity**. They can refer to a specific entity or not:
I'm going to buy a computer today.

Definite NPs

- the **definite** article (***the** book*),
- **demonstrative articles**
  (***this/that** book, **these/those** books*),
- **possessives** (***my/John's** book*)

Definite NPs can also consist of

- **personal pronouns** (***I, he***)
- **demonstrative pronouns** (***this, that, these, those***)
- universal **quantifiers** (***all, every)***
- (unmodified) **proper nouns** *(**John Smith, Mary, Urbana**)*

Definite NPs **refer to an identifiable entity**
(previously mentioned or not)

Information status

Every entity can be classified along two dimensions:

Hearer-new vs. hearer-old

Speaker assumes entity is (un)known to the hearer

Hearer-old: *I will call Sandra Thompson.*

Hearer-new: *I will call a colleague in California* (=Sandra Thompson)

Special case of hearer-old: **hearer-inferrable**

I went to the student union. The food court was really crowded.

Discourse-new vs. discourse-old:

Speaker introduces new entity into the discourse, or refers to an entity that has been previously introduced.

Discourse-old: *I will call her/Sandra now.*

Discourse-new: *I will call my friend Sandra now.*

Coreference resolution

Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old also became the Denver-based financial services company's president. It has been ten years since she came to Megabucks from rival Lotsabucks.

Coreference chains:

1. {Victoria Chen, Chief Financial Officer...since 2004, her, the 37-year-old, the Denver-based financial services company's president}
2. {Megabucks Banking Corp, Denver-based financial services company, Megabucks}
3. {her pay}
4. {rival Lotsabucks}

Coref as binary classification

Represent each NP-NP pair (+context) as a feature vector.

Training:
Learn a binary classifier to decide whether NP_i is a possible antecedent of NP_j

Decoding (running the system on new text):

- Pass through the text from beginning to end
- For each NP_i:
  Go through $NP_{i-1}...NP_1$ to find best antecedent NP_j.
  Corefer NP_i with NP_j.
  If the classifier can't identify an antecedent for NP_i, it's a new entity.

Features for Coref resolution

- Do the two NPs have the same **head noun**?
  (e.g. company)
- Do they contain the **same modifier**?
  (e.g. Denver-based)?
- Does the **gender** and **number** of the NPs match?
- Does one NP contain an alias (**acronym**) of the other?
  (United States = USA, Chief Executive Office = CEO)
- Is one NP a **hypernym/synonym** of the other?
- Is one NP an **appositive** of the other?
  [Victoria Chen], [CFO of Megabucks]
- Are both NPs **named entities** of the same type?

[CEO] = PERSON, Victoria Chen = PERSON

Evaluation: B-cubed F-score

The test data consists of D documents d with N total mentions m (mention boundaries are given as input)

- In the **gold standard**, each mention m belongs to a **'true' cluster** of mentions (=connected component) of size t_m
- In the **system output**, each mention m belongs to a **predicted cluster** of mentions (=connected component) of size p_m
- For each mention m, the **intersection** of the gold standard and system output clusters defines a **common cluster** of mentions of size c_m

$$\text{Precision } P = \frac{1}{N}\sum_{d\in D}\sum_{m\in d}\frac{c_m}{p_m}$$

$$\text{Recall } R = \frac{1}{N}\sum_{d\in D}\sum_{m\in d}\frac{c_m}{t_m}$$

$$\text{F-measure} = \frac{2PR}{P+R}$$

Special case: Pronoun resolution

Task: Find the antecedent of an anaphoric pronoun in context

*1. **John** saw **a beautiful Ford Falcon** at **the dealership**.*
*2. **He** showed **it** to **Bob**.*
*3. **He** bought **it**.*

he$_2$, it$_2$ = John, Ford Falcon, or dealership?
he$_3$, it$_2$ = John, Ford Falcon, dealership, or Bob?

Anaphoric pronouns

Anaphoric pronouns refer back to some previously introduced entity/discourse referent:
***John** showed **Bob** his car. **He** was impressed.*
*John showed Bob his car. **This** took five minutes.*

The **antecedent** of an anaphor is the previous expression that refers to the same entity.

There are number/gender/person **agreement constraints: *girls*** can't be the antecedent of ***he***
Usually, we need some form of **inference** to identify the antecedents.

Salience/Focus

Only some recently mentioned entities can be referred to by pronouns:

John went to Bob's party and parked next to a classic Ford Falcon.
He went inside and talked to Bob for more than an hour.
Bob told him that he recently got engaged.
He also said he bought it (???)/ the Falcon yesterday.

Key insight (also captured in Centering Theory)
Capturing which entities are salient (in focus) reduces the amount of search (inference) necessary to interpret pronouns!

Entity-based coherence

Discourse 1:
John went to his favorite music store to buy a piano.
It was a store John had frequented for many years.
He was excited that he could finally buy a piano.
It was closing just as John arrived.

Discourse 2:
John went to his favorite music store to buy a piano.
He had frequented the store for many years.
He was excited that he could finally buy a piano.
He arrived just as the store was closing for the day.

Entity-based coherence

Discourse 1:
John went to **his favorite music store** to buy **a piano**.
It was **a store John** had frequented for many years.
He was excited that **he** could finally buy **a piano**.
It was closing just as **John** arrived.

Discourse 2:
John went to **his favorite music store** to buy **a piano**.
He had frequented **the store** for many years.
He was excited that **he** could finally buy **a piano**.
He arrived just as **the store** was closing for the day.

How we refer to entities influences how coherent a discourse is (**Centering theory)**

Centering Theory

Grosz, Joshi, Weinstein (1986, 1995)

A linguistic theory of entity-based coherence and salience

- It predicts which entities are salient at any point during a discourse.
- It also predicts whether a discourse is entity-coherent, based on its referring expressions.

Centering is about **local (=within a discourse segment) coherence and salience**

Centering theory itself is **not a computational model** or an algorithm: many of its assumptions are not precise enough to be implemented directly. (Poesio et al. 2004)

- But many algorithms have been developed based on specific instantiations of the assumptions that Centering theory makes. The textbook presents a centering-based pronoun-resolution algorithm

Using Centering Theory for Summarization

Summarization

"The process of distilling the most important information from a text to produce and abridged version for a particular task and user"

- Abstract or extract?
- Generic (no specific task/user) or query-focused?
- Single-document or multi-document?

Output:

- Abstracts (of scientific papers)
- Headlines (or newspaper articles)
- Snippets (for webpages)
- Answers to complex questions (from multiple sources)

Extracts from a single document

Goal: Produce a paragraph that summarizes a document

1. Content selection:
   Find 'important' (key) sentences
   Extract key facts/phrases

2. Information ordering:
   What order should these key facts be presented in?

3. Sentence realization:
   Produce a coherent paragraph from the list of key facts

Centroid-based content selection

Which sentences are most **central** in a document?
Binary classification task: *sentence -> {include, don't include}*

Method A: Central sentences = **salient/informative** sentences:
- a **sentence is salient** if it contains many salient words.
- a **word is salient** (=informative) in a document
  if it occurs significantly more often than expected (if $-2 \log \lambda(w) > 10$)
  Likelihood ratio $\lambda(w)$: $P_{doc}(w)/P_{English}(w)$

Method B: Central sentences = most **similar** to other s's in doc.
- compute sentence-based TF/IDF for the words in a document
(sentence=TF/IDF's document, document= TF/IDF's collection)
- distance between sentences: cosine of TD/IDF vectors of all words
- centrality of sentence i: average distance to all other sentences in document

RST-based summarization

Use a **discourse parser** to identify rhetorical relations between sentences/clauses in a document.

This gives a **discourse tree** with hierarchical **nucleus-satellite** relations between clauses

This discourse tree defines a **salience ranking:**
the **highest nuclei** in the tree are the most salient

Information ordering and Sentence Realization

In which order should the key phrases be presented?

- Simplest case: order in which they appear in document
- Finding the optimal solution is NP-complete, but we can approximate

Use centering theory to measure **coherence**

- Use coreference resolution and parsing to produce an 'entity grid' (which entities occur in which sentence, and in which role), then find good sequences of transitions

Sentence realization may require some rephrasing:

- Use longer descriptions to introduce entities, shorter ones to refer back
  "**Bush** met Putin today. **George W. Bush** said…"
  => *"**George W. Bush** met Putin today. **Bush** said…"*

Happy fall break!