CS 4650/7650, Lecture 21
Discourse Structure

Jacob Eisenstein

November 12, 2013

As we have already seen, language has structured at multiple levels:

- The structure of words is **morphology**
- The structure of sentences is **syntax**
- The structure of sets of sentences is **discourse**.

Discourse relates to **dialogue structure**, which describes multi-party conversations, and **pragmatics**, which describes the role of context. It subsumes a number of related phenomena:

- identifying the characteristics that make text **coherent**. this is the analogue of **grammaticality**
- resolving references to entities and events
- describing rhetorical and narrative relationships between units of text.

1 Coherence

Meaningful discourses are **coherent**; random collections of grammatical clauses are not. The following examples are adapted from J&M:

1. Sue hid Lakshmi's keys. She was drunk.
2. ?? Sue hid Lakshmi's keys. She likes spinach.
3. Yangfeng offered spinach to anyone who could keep Lakshmi from driving. Sue hid Lakshmi's keys. She likes spinach.

Most people would find the first example reasonable, the second example to be strange at best, and the third example to be at least more reasonable than the second.

This suggests that readers have expectations about the relatedness of sentences in a discourse. In a discourse that describes a series of events, those expectations concern cause-and-effect relations. In a discourse that purports to be a well-reasoned argument, the expectations might be different.

1. You should wear sunscreen. It keeps your skin healthy.
2. ?? You should wear sunscreen. Exercise is a key part of a healthy lifestyle.

Rhetorical structure theory (RST) and the Penn Discourse Treebank attempt to capture these relationships.

Coherence is not just about the selection and order of clauses; it also relates to the choice of words within clauses. Grosz et al (1995) provide the following examples:

1. (a) **Serena** went to **her** favorite **music store** to buy a piano.
   (b) **It** was **a store** **Serena** had frequented for many years.
   (c) **She** was excited that **she** could finally buy a piano.
   (d) **It** was closing just as **Serena** arrived.
2. (a) **Serena** went to **her** favorite **music store** to buy a piano.
   (b) **She** had frequented **the store** for many years.
   (c) **She** was excited that **she** could finally buy a piano.
   (d) **She** arrived just as **the store** was closing for the day.

Which do you prefer? Many people find the second more coherent, because it is clearly about Serena, and features her in subject position (agent role) in each sentence. In the first version, the focus moves back and forth between Serena and the store. This sort of structure is captured by **centering** theory.

2 Applications of discourse

Why worry about discourse structure? It is connected to many NLP applications. As mentioned last time, discourse is intimately connected with reference resolution, but there are many other applications too.

2.1 Summarization

We would like to take long documents and compress them into short ones. Here is a minimal example from J&M:

- **First Union** Corp is continuing to wrestle with severe problems. According to industry insiders at Paine Webber, their **president**, **John R. Georgius**, **is planning to announce his retirement tomorrow**.
- First Union President John R. Georgius is planning to announce his retirement tomorrow.

We might also like to combine **multiple documents** into a single, short document: for example, to write the related work section of a research paper or project proposal.

2.2 Sentiment analysis

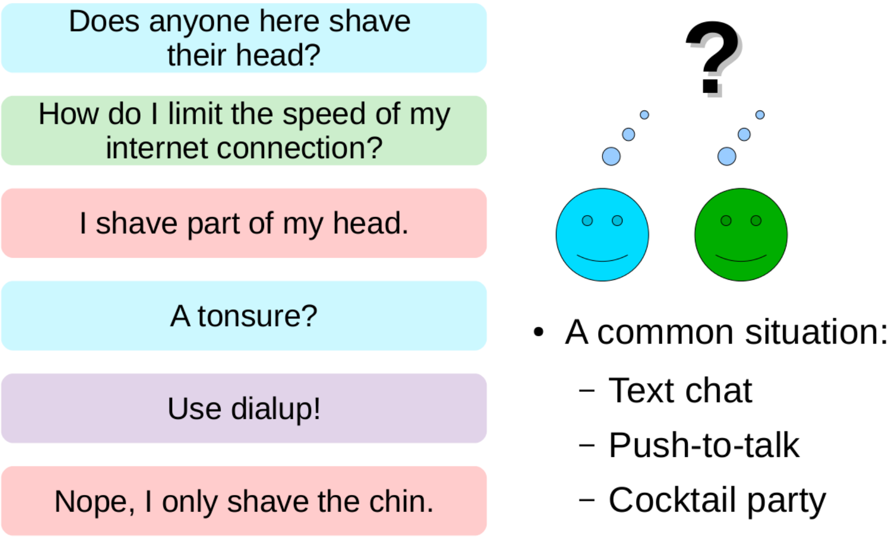
Bag-of-words approaches to sentiment analysis are hopelessly naive.

> *It could have been a* ***great*** *movie. It could have been* ***excellent****, and to all the people who have forgotten about the older,* ***greater*** *movies before it, will think that as well. It does have* ***beautiful*** *scenery, some of the* ***best*** *since Lord of the Rings. The acting is* ***well*** *done, and I really* ***liked*** *the son of the leader of the Samurai. He was a* ***likeable*** *chap, and I* ***hated*** *to see him die... But, other than all that, this movie is nothing more than hidden* ***rip-offs****.*

A model of discourse structure could reveal that the last sentence is the key to the entire paragraph.

2.3 Thread disentanglement

In many settings (e.g., chatrooms), we encounter language from multiple threads, which we must disentangle to understand. Figure from (Elsner and Charniak 2008).



Does anyone here shave their head?

How do I limit the speed of my internet connection?

I shave part of my head.

A tonsure?

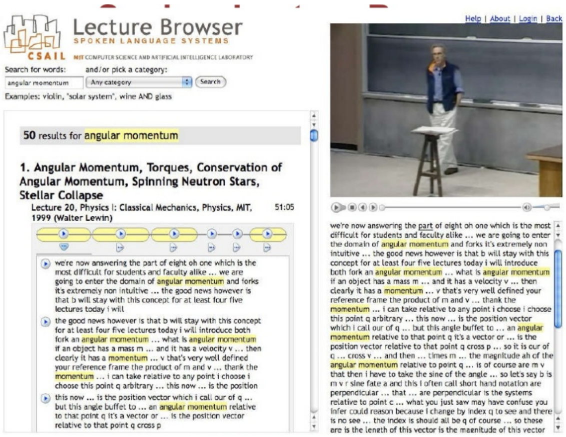
Use dialup!

Nope, I only shave the chin.

- A common situation:
  - Text chat
  - Push-to-talk
  - Cocktail party

2.4 Segmentation

Faced with long transcripts of text or speech, we would like to segment (possibly hierarchically) into chapters and sections.



Help | About | Login | Back
Lecture Browser
SPOKEN LANGUAGE SYSTEMS
CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY
Search

Search for words: angular momentum

and/or pick a category: Any category

Examples: violin, "solar system", wine AND glass

50 results for angular momentum

1. Angular Momentum, Torques, Conservation of Angular Momentum, Spinning Neutron Stars, Stellar Collapse

Lecture 20, Physics I: Classical Mechanics, Physics, MIT, 1999 (Walter Lewin) 51:05

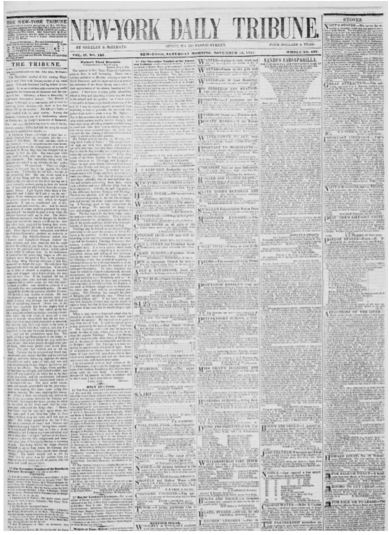
we're now answering the part of eight oh one which is the most difficult for students and faculty alike ... we are going to enter the domain of angular momentum and forks it's extremely non intuitive ... the good news however is that b will stay with this concept for at least four five lectures today i will

the good news however is that b will stay with this concept for at least four five lectures today i will introduce both fork an angular momentum ... what is angular momentum if an object has a mass m ... and it has a velocity v ... then clearly it has a momentum ... v that's very well defined your reference frame the product of m and v ... thank the momentum ... i can take relative to any point i choose i choose this point q arbitrary ... this now ... is the position

this now ... is the position vector which i call our of q ... but this angle buffet to ... an angular momentum relative to that point q it's a vector or ... is the position vector relative to that point q cross p

we're now answering the part of eight oh one which is the most difficult for students and faculty alike ... we are going to enter the domain of angular momentum and forks it's extremely non intuitive ... the good news however is that b will stay with this concept for at least four five lectures today i will introduce both fork an angular momentum ... what is angular momentum if an object has a mass m ... and it has a velocity v ... then clearly it has a momentum ... v that's very well defined your reference frame the product of m and v ... thank the momentum ... i can take relative to any point i choose i choose this point q arbitrary ... this now ... is the position vector which i call our of q ... but this angle buffet to ... an angular momentum relative to that point q it's a vector or ... is the position vector relative to that point q cross p ... so it is our of q ... cross v ... and then ... times m ... the magnitude ah of the angular momentum relative to point q ... is of course are m v that then i have to take the sine of the angle ... so let's say b is m v r sine fate a and this i often call short hand notation are perpendicular ... that ... are perpendicular is the systems relative to point c ... what you just saw may have confuse you infer could reason because i change by index q to see and there is no see ... the index is should all be q of course ... so these are is the length of this vector is the magnitude of this vector

This also arises in OCR of historical documents, where the original section formatting is often lost.



NEW-YORK DAILY TRIBUNE.

Yet another reason to do discourse segmentation is for **question answering**: paragraph-level answers can be obtained by matching automatically-obtained discourse segments with queries.

3 Discourse segmentation

Our goal in linear segmentation is to divide a text into **topically-coherent segments**. The theoretical foundation for linear segmentation is the notion of **cohesion**, introduced by Halliday and Hassan (1976).

There are several types of cohesion:

- **Reference**: use of pronouns to refer to entities defined elsewhere in the text (usually anaphora)
  *My brother is a great driver. **He** learned from the best.*

- **Substitution**: replacement of general phrases for more specific elements, when the meaning is clear
  *You should visit Pittsburgh, if you haven't already **done so**.*

- **Ellipsis**: the omission of words to avoid repeated phrases
  The younger child was outgoing, the older more reserved.

- **Conjunction** between phrases with *and*, *so*, *however*

- **Lexical**: repeated words, including synonyms etc

3.1 Coherence and cohesion

The key observation is that in coherent text, cohesive links occur within segments, not between them. But note that these concepts are different:

Cohesion without coherence

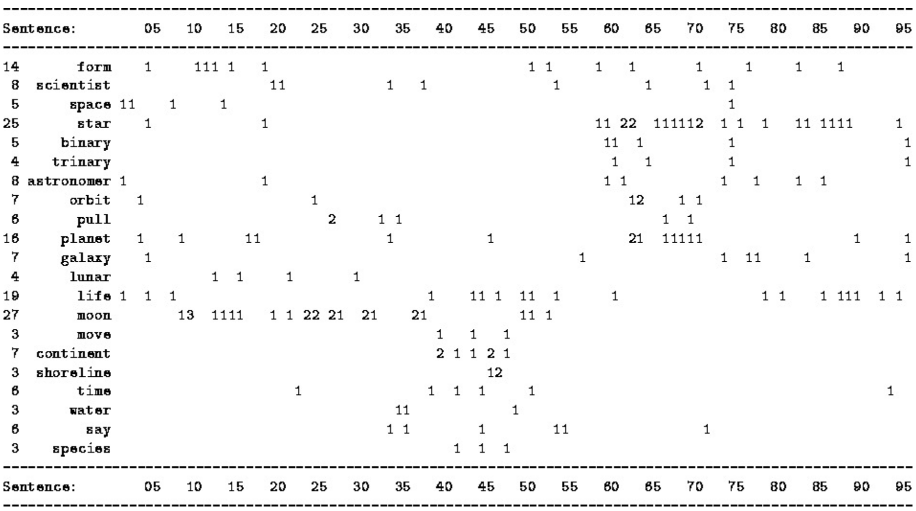
- *Wash and core six apples.*
- *Use the apples to cut out material for your new suit.*

Coherence without (lexical) cohesion

- *I came home from work at 6:00pm.*
- *Dinner consisted of two chicken breasts and a bowl of rice.*

Hearst (1993) showed that lexical cohesion could be quantified and used to support unsupervised topic segementation.

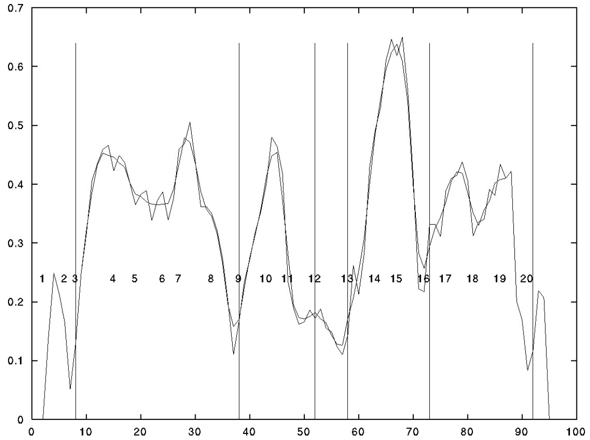
3.2 Lexical cohesion in topic segmentation



```
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
14       form    1     111 1   1                               1 1     1   1       1     1     1    1
 8  scientist                   11            1   1               1          1      1  1
 5      space 11    1     1                                                            1
25       star    1             1                                       11 22  111112  1 1  1   11 1111     1
 5     binary                                                           11  1          1                    1
 4    trinary                                                            1   1         1                    1
 8 astronomer 1                1                                        1 1           1   1    1  1
 7      orbit   1                    1                                     12    1 1
 6       pull                          2     1 1                               1  1
16     planet   1    1       11               1           1                21  11111                  1     1
 7     galaxy    1                                                   1                1  11     1           1
 4      lunar            1  1     1       1
19       life 1  1  1                              1    11 1  11  1      1                 1 1    1 111  1 1
27       moon        13  1111   1 1 22 21  21    21           11 1
 3       move                                       1   1   1
 7  continent                                       2 1 1 2 1
 3  shoreline                                             12
 6       time                     1               1  1  1     1                                          1
 3      water                                  11            1
 6        say                                 1 1        1        11                1
 3    species                                         1  1  1
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
```

TextTiling uses smoothed cosine similarity between adjacent spans of text, putting segmentation boundaries at the troughs:

$$\mathrm{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$



We can also measure lexical cohesion probabilistically.

- Each segment has a distribution over words $\boldsymbol{\theta}$,

$$\boldsymbol{x}|z, \boldsymbol{\theta} \sim \text{Multinomial}(\theta_z) \tag{1}$$

- Choose the segmentation and distributions that maximize the marginal likelihood (Utiyama and Isahara, 2001)

$$\hat{\boldsymbol{z}}, \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{z},\boldsymbol{\theta}} P(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \prod_n P(\boldsymbol{x}_n | \boldsymbol{\theta}_{z_n}) \tag{2}$$

- This is like an HMM, but with a **left-to-right** constraint: we can never revisit a state once we have left it.

- But the distributions are a nuisance parameter. Let's marginalize them (Eisenstein and Barzilay 2008):

$$\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \int_{\boldsymbol{\theta}} P(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\theta}) P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}) \prod_n P(\boldsymbol{x}_n | \boldsymbol{\theta}_{z_n}) \tag{3}$$

- Why a probabilistic framework? Because we can also incorporate discourse **connectors**: words or phrases that tend to appear at the start of segments. More on this later.

4 Centering

As we saw on Tuesday, there are deep relationships between how a noun phrase is mentioned, and the structure of the discourse.

- On tuesday, we considered the form of referring expressions: fully-specified name (*Angela Merkel*), partially-specified name (*Merkel*), definite NP (*the German Chancellor*), indefinite NP (*a German politician*), pronoun (*she*)
- Centering theory focuses on the syntactic position of each entity: subject, object, or oblique. This was motivated by the example earlier.
- One explanation is that salience reduces the amount of search necessary to interpret pronouns.

At each utterance U_n, we have:

- A backward-looking center $C_b(U_n)$:
  the entity currently **in focus** after U_n.
- A forward-looking center $C_f(U_n)$:
  an ordered list of candidates for $C_b(U_{n+1})$.
- The top choice in $C_f(U_n)$ is $C_p(U_{n+1})$

How do we order the candidates from $C_b(U_{n+1})$ to the forward-looking center? By syntax:

1. Subject
   Abigail *saw an elephant.*
2. Existential predicate nominal
   There is ***an elephant*** *in the room.*
3. Direct object
   Abigail gave ***a snack*** *to the elephant.*
4. Indirect object or oblique
   Abigail gave a snack to ***the elephant****.*
5. demarcated adverbial prepositional phrase
   Inside ***the zoo****, the elephant is king.*

Rule: If any element of $C_f(U_n)$ is realized by a pronoun in U_{n+1}, then $C_b(U_{n+1})$ must also be realized as a pronoun.

- Generate possible C_b and C_f for each set of reference assignments
- Filter by constraints: syntax, semantics, and centering rules

- Rank by transition orderings: continue, retain, smooth-shift, rough-shift

	$C_b(U_{n+1}) = C_b(U_n)$ or $C_b(U_n) = \varnothing$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-shift

In a coherent discourse, we select transitions according to the following preferences: continue, retain, smooth-shift, rough-shift

4.1 Example

Here's an example of how to use centering to resolve pronouns.

U_n	$C_f(U_n)$	$C_p(U_n)$	$C_b(U_n)$	transition
John saw a beautiful Masi at the bike shop	John, Ford, bike shop	John	$\varnothing$	
He showed it to Bob	John, Masi, Bob	John	John	Continue
He showed it to Bob	John, bike shop, Bob	John	John	Continue
He bought it	John, Masi or bike shop	John	John	Continue
He bought it	Bob, Masi or bike shop	Bob	Bob	Smooth-shift

- Centering theory tells us that we prefer *John* over *Bob* as the referent for *he* in U_3, because this would be a continue transition rather than a smooth-shift.

- Centering doesn't really give us a rule for choosing *Masi* over *bike shop* in U_2, because neither is $C_b(U_2)$. We might apply the grammatical role hierarchy since there is no other basis for this decision.

4.2 Centering from entities

- Barzilay and Lapata (2005, 2008) showed how to build computational models of centering and local coherence, using shallow NLP.

- Their system did not fully solve the entity coreference problem, but was still accurate enough to support useful applications.

- Key idea: represent centering through an **entity grid**.

Each grid is a vector $\boldsymbol{x}$; we can compute a probability distribution $P(\boldsymbol{x}) = \prod_n P(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})$. This probability distribution captures preferences for various centering transition.

- Elsner and Charniak (2011) showed how to use entity grids to do thread disentanglement.
- The entity grid can also be used to measure the coherence of text. This idea is used in the automated essay scoring for the GRE (Burstein et al 2010).

5 Rhetorical structure theory

RST (Mann & Thompson, 1987) describes coherence relations between **utterances**

- Utterances are not exactly sentences and not exactly clauses...
- More technically: *elementary discourse units* (EDUs).
  Identifying them is a preprocessing step (this is true for centering as well).

Utterances are connected by rhetorical relations.

- Asymmetric relations have a nucleus and a satellite.
- Symmetric relations have multiple nucleii.
- Some relations:
  - **Elaboration**: *Yangfeng hid the keys, placing them behind a plant.*
  - **Attribution**: *Sara claimed that Twitter would prove to be a good investment.*
  - **Evidence**: *Leslie must like bikes. She owns five of them.*
  - **Contrast**: *Abed likes science fiction, but Annie thinks it's silly.*
  - **Concession**: *Abed likes science fiction, even though he knows it's silly.*
  - **List**: *Jeff is the leader; Britta is the political activist; Troy is comic relief.* Note that List is a multi-nuclear relation.

These relations can be composed into a hierarchical structure.

5.1 Discourse parsing

RST parsing is the task of recovering this structure from a text. There are two substeps:

1. **Segmentation** of text (sentences) into elementary discourse units (EDUs). Soricut and Marcu present a simple generative approach:

   $$P(b|\text{words}, \text{tree}) = \frac{\text{count}(N_p \to \ldots N_w \uparrow N_r \ldots)}{\text{count}(N_p \to \ldots N_w N_r \ldots)}$$

   They use lexicalized non-terminals, e.g., $P(b|VP(\textit{says}) \to VBZ(\textit{says})S(\textit{will}))$

2. **Identifying** the tree of relations in the discourse. In a simple model, we might represent the probability of a tree as the product of the probabilities of the relations, like a labeled dependency parse. These probabilities could be based on features of the EDUs.

Sagae (2009) presents a shift-reduce approach to discourse parsing, which is good for speed.

- **shift**: push next token of input onto the stack
- **reduce-left-LABEL**
  - pop top two elements of stack,
  - create a relationship of type LABEL between them.
  - headed by element on left
- **reduce-right-LABEL**: same, but headed by the element to the right

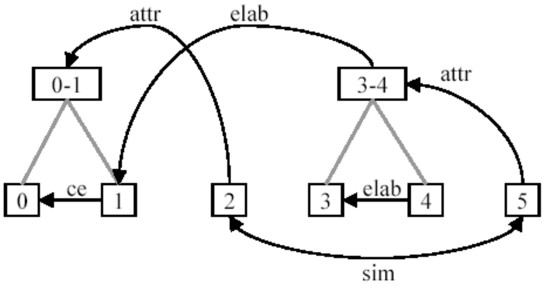
The choice of which move to make is the output of an averaged perceptron. Let's apply this to an example.

5.2 Discourse graphs

Wolf and Gibson (2003) argue that many discourses cannot be fully described by a tree, and that general graphs are more appropriate. Here's an example:

0. Farm prices in October edged up 0.7% from September
1. as raw milk prices continued to rise,
2. the Agriculture Department said.

3. Milk sold to the nations dairy plants and dealers averaged $14.50 for each hundred pounds,
4. up 50 percent from September and up $1.50 from October 1988,
5. the department said.



6 Discourse relations

- Rhetorical Structure Theory (RST) builds a **tree** that completely covers all the text.
- The Penn Discourse Treebank (PDTB) takes a different approach.
  - Spans of text are related by discourse connectors (e.g., *however*), which may be explicit or implicit.
  - Not every span may participate in a relation.
  - Some spans may participate in multiple relations.
  - This is sometimes called "discourse chunking."

See slides for examples.

7 Essay scoring

Here is the first paragraph from an essay that scores a 5 on the GRE.

1. *In today's society, college is ambiguous.*
2. *We need it to live, but we also need it to love.*
3. *Moreover, without college most of the world's learning would be egregious.*

4. *College, however, has myriad costs.*

5. *One of the most important issues facing the world is how to reduce college costs.*

6. *Some have argued that college costs are due to the luxuries students now expect.*

7. *Others have argued that the costs are a result of athletics.*

8. *In reality, high college costs are the result of excessive pay for teaching assistants.*[1]

Based on what you know about discourse, why do you think it scores a 5?

8 Other discourse structures

- **Conventionalized structures**. In some genres, texts have tightly prescribed structures. For example, research papers in NLP begin with an abstract, followed by an introduction, related work, model, implementation, experiment, discussion. In Wikipedia, articles about politicians, athletes, movies, and nations have highly conventionalized structures too. This is closely related to work on **functional zoning**.

- **Events and states**. Some documents describe processes (for example, medical procedures or chemical reactions; see (Scaria 2013) at EMNLP), or scripts (for example, instructions on how to cash a check). In this case, the relevant structure are the steps of the process, the state of the entities involved, and the **constraints** between these events and states.

- **Intentions**. Suppose we are trying to get something to happen: for example, to get someone to believe a claim, or to know how to perform an action. We can view discourse as a **planning** problem: choosing the right sequence of utterances to achieve our end goal, given a model of the preconditions and results of each utterance.

- **Plot and narrative**. In fictional documents, the relevant structure is the plot itself. By learning a representation of the plot, we can

[1] From Lee Perelman, `http://www.cbc.ca/spark/wp-content/uploads/2012/05/Essays-for-Robo-Reader.pdf`

understand character's motivations and build a model of suspense. Mark Riedl here at GT is an expert on this.