CS689: Computational Linguistics for Indian Languages
Discourse

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs689/

2nd semester, 2023-24

Tue 10:30–11:45, Thu 12:00–13:15 at RM101/KD102

Discourse

- Discourse is consistent and coordinated text
- Coherence is relationship between *sentences*
- Relations are called coherence relations or discourse relations
- “She took a train from Kanpur to Kolkata. She likes potatoes.”
- “She took a train from Kanpur to Kolkata. She had to attend a seminar.”
- “She took a train from Kanpur to Kolkata. She likes fish.”
- Local coherence is within sentences that are adjacent or close by
- Global coherence is overall coherence in the entire text
- Essentially, what the text is about

Coherence Models

- About *someone* or *something*
- Salient entity
- Entity-based coherence model
- Incoherent set of sentences jump across entities
    - "Ram went to the market. Gita was playing in her room. He bought vegetables. She fell asleep."
- Requires coreference resolution
- Nearby sentences are about the same *topic*
- Topical coherence model
- Entities are from the same *semantic field* or *topic*
- Thus, related words are used
- Lexical cohesion
    - "India has always produced good batsmen. Nowadays, bowling has become fearsome as well. Possibly, the improvement of domestic tournaments helped."

Coreference

- Coreferences are mentions or referring expressions about the same referrent
- *Referring expressions* to the same *discourse entity* corefer
- A discourse model is said to be constructed in the mind of reader/listener
- An entity is evoked into the model in its first mention
- Later mentions access the representation

Example

- "Yuvraj, the stylish left-handed batsman and captain of Delhi Capitals, saw his pay rise to 20 crores. Thus, he doubled his pay by switching over to Delhi from Mumbai."
- Coreference chains or coreference clusters
  - { Yuvraj, the stylish left-handed batsman, captain of Delhi Capitals, his, he }
  - { his pay, 20 crores }
  - { Delhi Capitals, Delhi }
  - { Mumbai }
- Coreference resolution is the task of finding these chains
- Entity linking or entity resolution is mapping each discourse entity to a real-world entity
- Coreferences can be to other types
  - Events: "The switch raised eyebrows."
  - Discourse segments: "But that turned out to be a lie."

Anaphora and Cataphora

- When an entity is first introduced and then referred, this reference is called anaphora
- Referring expression is called anaphor or anaphoric
- Anaphor corefers with a prior mention, which is called antecedent
  - Even before Damayanti married Nala, she knew about him
- When a mention is introduced before the entity, it is called cataphora
- Referring expression is called cataphor or cataphoric
- Cataphor corefers with a prior mention, which is called precedent
  - Even before she married him, Damayanti knew about Nala

Linguistic Clues

- Number agreement
  - "The farmers surrounded the leader. He was questioned."
  - "IIT Kanpur launched a new website. They worked on it for months."
- Person agreement
  - "I used to think of him as my best friend."
  - "I supported him because of my belief", she said.
- Gender agreement
  - "Gita designed a new algorithm. It is so exciting."
  - "Gita designed a new algorithm. She is so exciting."
- Type of pronoun: *reflexive*, etc.
  - "Sita bought herself a cycle."
  - "Sita bought her a cycle."

Other Clues

- Recency
    - "Gita owns a 30-year old cassette player. Sita owns an even older one. It can still play songs."
- Grammatical role
    - "Rama went to the battle with Lakshmana. He killed Ravana."
    - "Vibhisana took Lakshmana to the room. He killed Indrajit."
- Verb semantic role emphasis
    - "Ram called Shyam. He has lost the keys."
    - "Ram scolded Shyam. He has lost the keys."
- Semantic restrictions
    - "She ate the food in her new plate after cooking it."
    - "She ate the food in her new plate after washing it."
    - "She ate the food in her new plate after heating it."

Difficulties in Coference Resolution

- Absence of entities
  - “He doesn’t own a car. *It is red.”
- Generic
  - “I love mangoes. They are tasty.”
- Bounding
  - “When the workers were asked, each one raised his hand.”
- World knowledge and beliefs
  - “The police denied the demonstrators a permit because they feared violence.”
  - “The police denied the demonstrators a permit because they advocated violence.”
- Stereotypes
  - “The head summoned both the doctor and the nurse, but especially scolded him for the event.”
  - “The head summoned both the doctor and the nurse, but especially scolded her for the event.”

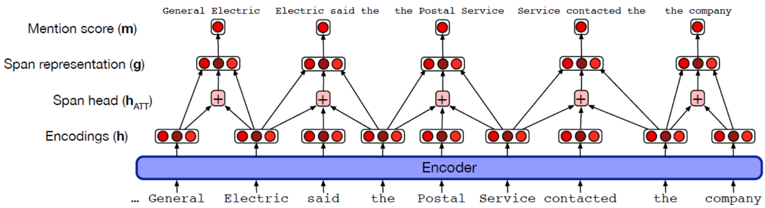
Mention Detection

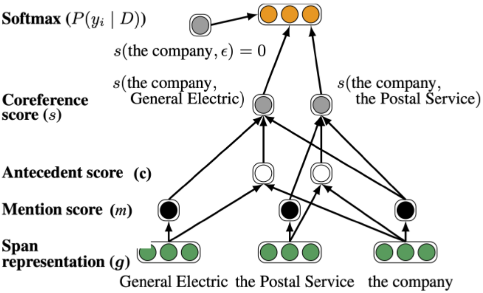
- Mentions are *referring expressions*
- Mention detection aims to find spans of text that are *mentions*
- Typically, all *noun phrases*, *pronouns* and *named entities*
- High recall but low precision
- Can be rule-based
- Neural network based
- Can add another classifier to determine whether *referring (anaphoric/cataphoric)* or *referred (antecedent/precedent)*

Mention-based Architectures

- Mention-Pair architecture
- Given a candidate pair of mentions, a *referring* and a *referrent*, is it a *coreference pair*?
  - Yes/no binary classifier
- Mention-Rank architecture
- Given a candidate *referrent*, *rank* all candidate *referring* mentions
  - Find probability of each referring mention, including *null*
  - Choose highest
- *Entity-based* models work with *discourse entities* instead of mentions
- Classifiers use a lot of features
  - Position of word, Head word, POS tag, morphological tag, word distance, etc.

Neural Network Architectures




Softmax ($P(y_i \mid D)$)
$s(\text{the company}, \epsilon) = 0$
s(the company, General Electric)
s(the company, the Postal Service)
Coreference score (s)
Antecedent score (**c**)
Mention score (m)
Span representation (g)
General Electric
the Postal Service
the company

- Dataset is augmented using non-maximal coreferences as negative examples

Rhetorical Structure Theory

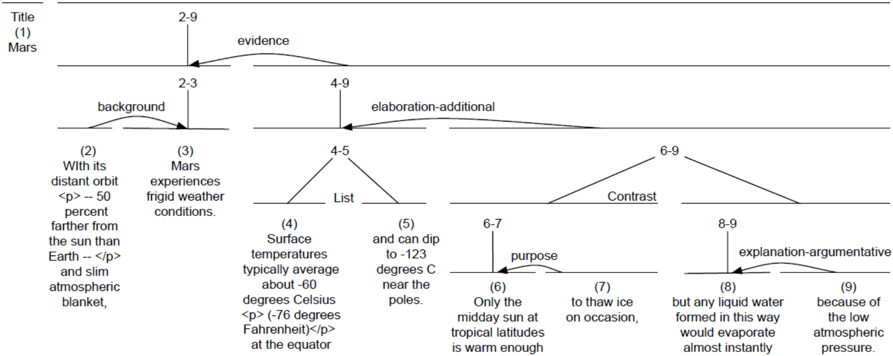
- Rhetorical Structure Theory (RST) is a model for discourse organization
- RST relations are between two spans of text, nucleus and satellite
- *Nucleus* is more central
- *Satellite* is interpretable with respect to nucleus
- Text is broken into spans that have RST coherence relations between them
- Each text span is called an elementary discourse unit (EDU) or discourse segment
- Finding such EDUs and their relations is the discourse resolution task
    - Similar to *dependency parsing*

RST Coherence Relations

- Mostly, relations are asymmetric (from satellite to nucleus)
- Reason: Satellite gives reason for action of nucleus
  - "[She took a train from Kanpur to Kolkata.]$_{\text{NUC}}$ [She had to attend a seminar.]$_{\text{SAT}}$"
- Elaboration: Satellite gives additional information
  - "[She was from Sundarbans.]$_{\text{NUC}}$ [She lived in the midst of dense jungles.]$_{\text{SAT}}$"
- Evidence: Satellite gives evidence or support
  - "[She must be in the building.]$_{\text{NUC}}$ [Her car is parked outside.]$_{\text{SAT}}$"
- Attribution: Satellite gives source of attribution
  - "[According to experts,]$_{\text{SAT}}$ [Europe is going into recession.]$_{\text{NUC}}$"

RST Example

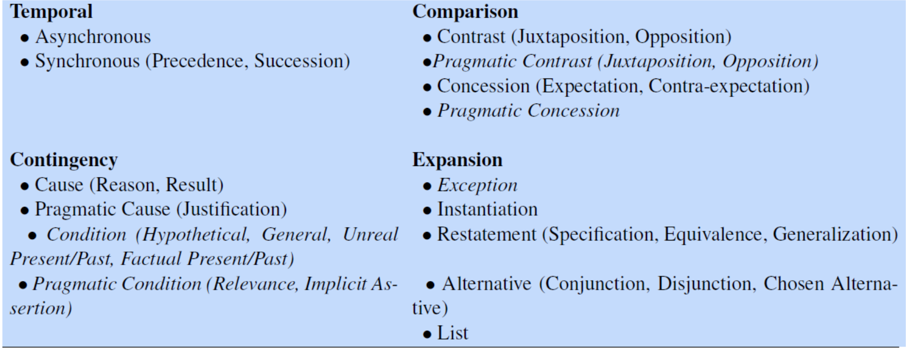
- With its distant orbit–50 percent farther from the sun than Earth–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator and can dip to -123 degrees C near the poles. Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure.



Title (1) Mars
2-9
evidence
2-3
background
4-9
elaboration-additional
(2) WIth its distant orbit <p> -- 50 percent farther from the sun than Earth -- </p> and slim atmospheric blanket,
(3) Mars experiences frigid weather conditions.
4-5
List
6-9
Contrast
(4) Surface temperatures typically average about -60 degrees Celsius <p> (-76 degrees Fahrenheit)</p> at the equator
(5) and can dip to -123 degrees C near the poles.
6-7
purpose
8-9
explanation-argumentative
(6) Only the midday sun at tropical latitudes is warm enough
(7) to thaw ice on occasion,
(8) but any liquid water formed in this way would evaporate almost instantly
(9) because of the low atmospheric pressure.

Penn Discourse TreeBank

- Penn Discourse TreeBank (PDTB) discourse relations
- Discourse connectives signal discourse relations
- Specific words or word sequences
  - because, although, when, as a result
- Sentence-to-sentence connections
- Flat structure



Temporal

- Asynchronous
- Synchronous (Precedence, Succession)

Comparison

- Contrast (Juxtaposition, Opposition)
- *Pragmatic Contrast (Juxtaposition, Opposition)*
- Concession (Expectation, Contra-expectation)
- *Pragmatic Concession*

Contingency

- Cause (Reason, Result)
- Pragmatic Cause (Justification)
- *Condition (Hypothetical, General, Unreal Present/Past, Factual Present/Past)*
- *Pragmatic Condition (Relevance, Implicit Assertion)*

Expansion

- *Exception*
- Instantiation
- Restatement (Specification, Equivalence, Generalization)
- Alternative (Conjunction, Disjunction, Chosen Alternative)
- List

Discourse Levels: Saṅgati

- Discourse relations are called saṅgati (सङ्गति)
- Saṅgati expresses continuity and proper positioning of text
- Induces the desire to know what is being said next in text
- Saṅgati relations can be at different levels
- Śāstra saṅgati (शास्त्र सङ्गति): coherence at the level of a *subject* of a book
- Adhyāya saṅgati (अध्याय सङ्गति): coherence at the level of a *chapter* of a book
- Pāda saṅgati (पाद् सङ्गति): coherence at the level of a *section* of a chapter
- Adhikaraṇa saṅgati (अधिकरण सङ्गति): coherence at the level of a *topic*

Topic Levels

- Topic-level analysis can be, again, at multiple levels
- *Sentential analysis*: Establishing relations among words in a sentence
- *Paragraph level analysis*: Identifying inter-sentential relations based on either explicit or implicit connectives
- *Sub-topic level analysis*: Establishing relations between successive paragraphs showing the consistency of the argument leading to a sub-topic
- *Topic level analysis*: Showing relevance of each sub-topic towards the goal of the main topic and, thus, the coherence

Topic or Adhikaraṇa

- Adhikaraṇa expresses a topic
- Adhikaraṇa typically consists of 5 aspects
  - Viṣaya (विषय): Topic
  - Saṃśaya (संशय): Doubts
  - Vicāra (विचार): Discussions or Arguments
  - Nirṇaya (निर्णय): Conclusion
  - Prayojanam (प्रयोजनम्): Purpose

Topic-level or Adhikaraṇa Saṅgati

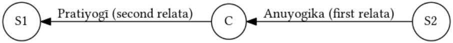
- Adhikaraṇa expresses a topic
- Topic-level or Adhikaraṇa saṅgati are 6
- Prasaṅga (प्रसङ्ग): Corollary or Related
- Upodghāta (उपोद्घात): Pre-requisite or Introduction
- Hetutā (हेतुता): Causal dependence
- Avasara (अवसर): Opportunity for further enquiry
- Nirvāhakaikya (निर्वाहकैक्य): Having same cause
- Kāryaikya (कार्यैक्य): Having same effect

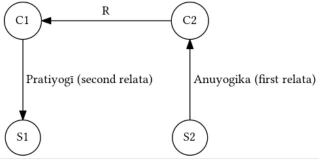
Sub-topic level Saṅgati

- Saṅgati that binds one adhikaraṇa to another
- Sub-topic level analysis
- Different schools of philosophy use different subsets
- Praśna (प्रश्न): Question
- Uttara (उत्तर): Answer
- Vyākhyā (व्याख्या): Elaboration
- Ākṣepa (आक्षेप): Objection
- Samādhāna (समाधान): Justification
- Dṛṣṭānta (दृष्टान्त): Example
- Pratyudāharaṇa (प्रत्युदाहरण): Counter-example
- Apavāda (अपवाद): Exception
- Prasaṅga (प्रसङ्ग): Corollary / Incidental illustration
- Upodghāta / Utpatti (उपोद्घात / उत्पत्ति): Pre-requisite / Introduction
- Bādhaka (बाधक): Rejection
- Sādhaka (साधक): Reaffirmation
- Dūṣaṇa (दूषण): Criticism

Inter-Sentential Relations

- Two arguments (sentences) of an inter-sentential connective are `anuyogika` and `pratiyogī`
- Connectives can be *single* or *parallel*





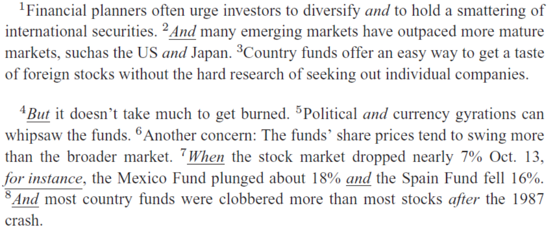
- Relations can be *link words* between two verbs
  - `kintu parantu atha apica`
  - `yadi-tarhi yadyapi-tathāpi yadā-tadā`
- Relations can be *non-finite verbs*
  - `ktvā`: activity preceding the main activity
  - `tumun`: purpose of the main activity
  - `śatṛ śānac`: simultaneity of two activities
- Resembles PDTB structure

Types of Relations

- Samānakālaḥ (simultaneity of time): yadā, tadā, yasmin kāle, tasmin kāle
- Samānādhikaraṇatvam (colocation): yatra, tatra, yasmin, tasmin
- Hetuhetumadbhāvaḥ (cause effect relationship): yataḥ, tataḥ, yasmāt, tasmāt, ataḥ
- Asāphalyam (failure): kintu
- Anantarakālīnatvam (following action): atha
- Kāraṇasatve'api kāryābhāvaḥ / kāraṇābhāve'api kāryotpattiḥ (non-productive effort or product without cause): yadyapi, tathāpi, athāpi
- Pratibandhaḥ (conditional): yadi, tarhi, cet, tarhyeva
- Samuccayaḥ (conjunction): ca, apica, kiñca
- Pūrvakālīkatvam (precedence): ktvā
- Samānakālīkatvam (Simultaneity of two activities): śatṛ, śānac
- Prayojanam (purpose of the main activity): tumun
- Implicit (invisible) relations

Shallow Parsing

- Shallow discourse parsing to find out PDTB structure
- Find discourse connectives (disambiguated from non-discourse uses)
- Find spans for each connective
- Label relationship between spans
- Assign relation between pair of sentences (if not already a span)



[1]Financial planners often urge investors to diversify *and* to hold a smattering of international securities. [2]*And* many emerging markets have outpaced more mature markets, suchas the US *and* Japan. [3]Country funds offer an easy way to get a taste of foreign stocks without the hard research of seeking out individual companies.

[4]*But* it doesn't take much to get burned. [5]Political *and* currency gyrations can whipsaw the funds. [6]Another concern: The funds' share prices tend to swing more than the broader market. [7]*When* the stock market dropped nearly 7% Oct. 13, *for instance*, the Mexico Fund plunged about 18% *and* the Spain Fund fell 16%. [8]*And* most country funds were clobbered more than most stocks *after* the 1987 crash.

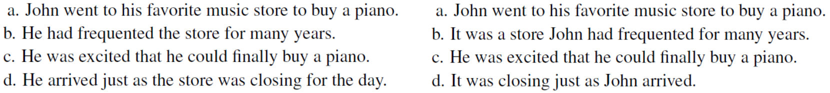
- Example
  - Candidate discourse words are italicized while actual are underlined
  - "When": "the … Oct 13" and "for instance … 16%"
  - Relation sense is Synchrony
  - Sentence 1 – Sentence 2: explicit Expansion.Conjunction
  - Sentence 5 – Sentence 6: implicit AltLex

Algorithm

- Machine learning setup
- Features of words, etc.
- POS tags
- Parse features
- Features of words around it
- Rules
- Position rules

Centering Theory

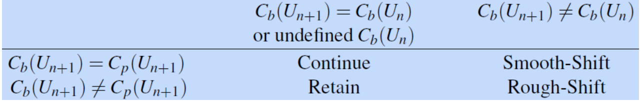
- Centering Theory is an entity-based discourse theory



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

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

- Discourse in which adjacent sentences maintain the same "centered" (or salient) entity is more coherent
- For a sentence or utterance U_n, two centers
- Backward-looking center $C_b(U_n)$: current salient entity (after U_n)
- Forward-looking centers $C_f(U_n)$: *set* of potential future salient entities (potential $C_b(U_{n+1})$)
  - $C_p(U_n)$ is *most preferred* future center
- *Centering transitions* based on U_n and U_{n+1}

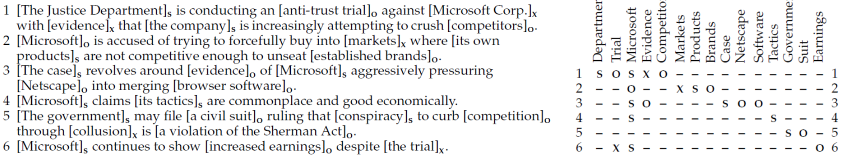


	$C_b(U_{n+1}) = C_b(U_n)$ or undefined $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

- Continue $\succ$ Retain $\succ$ Smooth-Shift $\succ$ Rough-Shift

Entity Grid Model

- Entity Grid Model tracks discourse entities
- In a sentence, a discourse entity can have 4 roles
    - Subject (S), Object (O), Neither (X), Absent (–)



1 [The Justice Department]$_s$ is conducting an [anti-trust trial]$_o$ against [Microsoft Corp.]$_x$ with [evidence]$_x$ that [the company]$_s$ is increasingly attempting to crush [competitors]$_o$.
2 [Microsoft]$_o$ is accused of trying to forcefully buy into [markets]$_x$ where [its own products]$_s$ are not competitive enough to unseat [established brands]$_o$.
3 [The case]$_s$ revolves around [evidence]$_o$ of [Microsoft]$_s$ aggressively pressuring [Netscape]$_o$ into merging [browser software]$_o$.
4 [Microsoft]$_s$ claims [its tactics]$_s$ are commonplace and good economically.
5 [The government]$_s$ may file [a civil suit]$_o$ ruling that [conspiracy]$_s$ to curb [competition]$_o$ through [collusion]$_x$ is [a violation of the Sherman Act]$_o$.
6 [Microsoft]$_s$ continues to show [increased earnings]$_o$ despite [the trial]$_x$.

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	s	o	s	x	o	–	–	–	–	–	–	–	–	–	–	1
2	–	–	o	–	–	x	s	o	–	–	–	–	–	–	–	2
3	–	–	s	o	–	–	–	–	s	o	o	–	–	–	–	3
4	–	–	s	–	–	–	–	–	–	–	–	s	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	s	o	–	5
6	–	x	s	–	–	–	–	–	–	–	–	–	–	–	o	6

- Patterns of local entity transition
    - For "department", the first (binary) pattern is `S–`
- Each transition has an overall *global* probability
- Each document has a probability distribution of transitions
- Can be used as machine learning features to predict *human discourse scores*

Evaluating Discourse

- Human ratings are the best
  - Not scalable
- Assume that a human-produced document has the “best” discourse
  - At least, better than *pseudo-documents* that contain random permutations of same set of sentences
- 3 tasks
- Sentence order discrimination task
  - For every document, produce *n* permutations of its sentences
  - Discourse coherence algorithm should rank original document as best
- Sentence insertion task
  - Take out a random sentence out of *n*
  - Create copies by inserting it at all possible places
  - Harder since order of most sentences remain the same
- Sentence order reconstruction task
  - Randomize the order of sentences
  - Task is to produce the correct order
  - Hardest