INTRODUCTION TO DISCOURSE LINGUISTICS AND DISCOURSE STRUCTURE

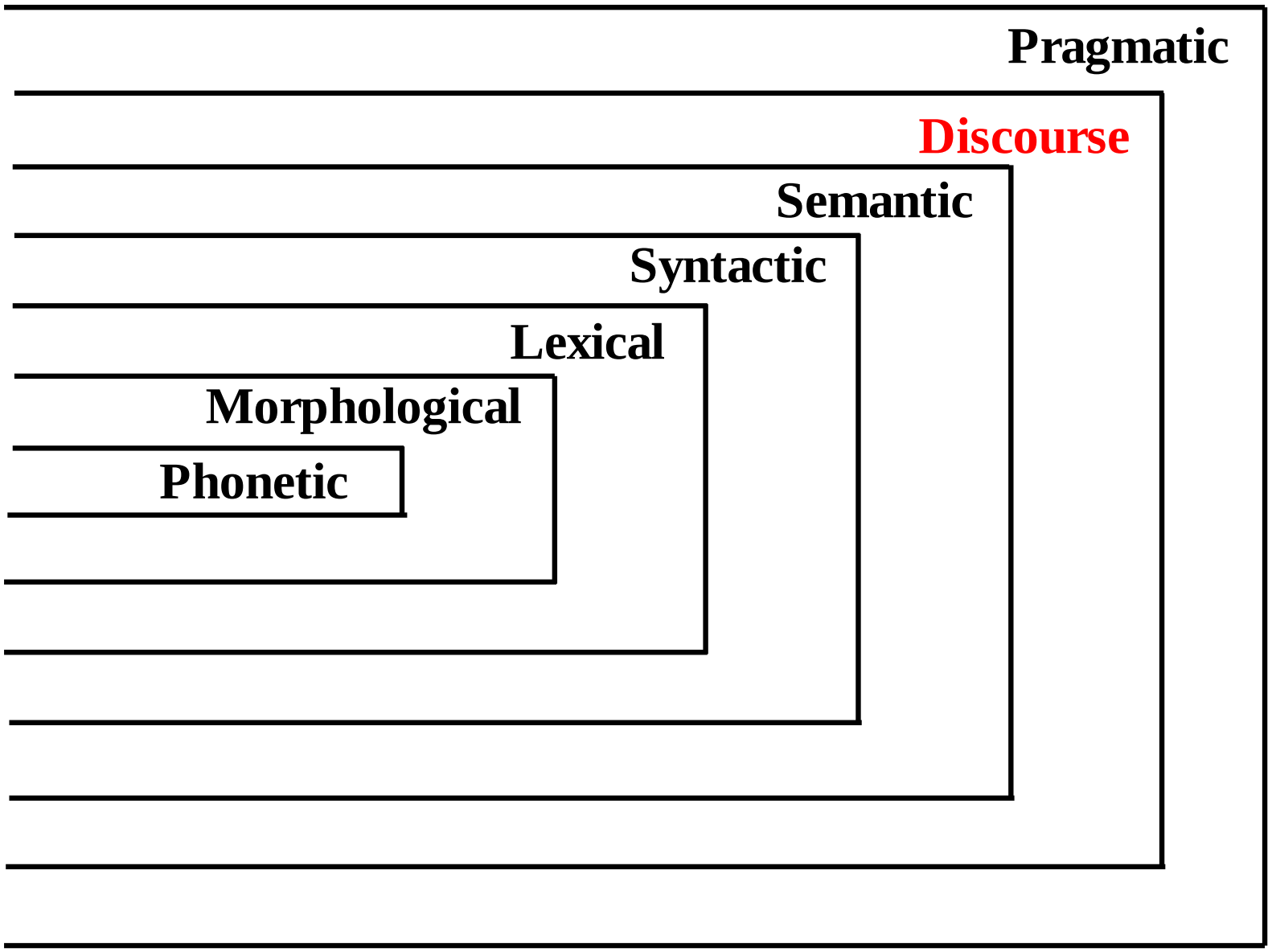
Lu Xiao
lxiao04@syr.edu
213 Hinds Hall

adopted some materials developed in previous courses by Nancy McCracken, Liz Liddy and others; and some instructor resources for the book "Speech and Language Processing" by Daniel Jurafsky and James H. Martin

HOMEWORK 4

- Choose ONE data file to work on
- Lab 10 and 11 may be useful

SYNCHRONIC MODEL OF LANGUAGE



DISCOURSE LINGUISTICS

" *No one is in a position to write a comprehensive account of discourse analysis. The subject is at once too vast, and too lacking in focus and consensus*." (Stubbs, Discourse Analysis)

Definitional Elements

- Study of texts (linguistic units) larger than a sentence.
- Text is more than a sequence of sentences to be considered one by one.
  - Rather, sentences of a text are elements whose significance resides in the contribution they make to the development of a larger whole.
- Each type of text has its own structure that can convey meaning to the reader.
- Some issues of discourse understanding are closely related to those in pragmatics which studies the real world dependencies of utterances.

Distinctions Between Text And Discourse

- In some contexts, e.g. in communication research, the word discourse means
  - interactive conversation
  - spoken
- And the word text means
  - non-interactive monologue
  - Written
- But for (American) linguists, the word discourse can mean both of these things at the discourse level.

Scope of Discourse Analysis

- What does discourse analysis extract from text more than the explicit information discoverable by sentence-level syntax and semantics methodologies?
  - Structural organization of the text
  - Overall topic(s) of the text
  - Features which provide *cohesion* to the text
- What linguistic features of texts reveal this information to the analyst?

Discourse Segmentation

- Documents are automatically separated into passages, sometimes called fragments, which are different discourse segments
  - Discourse segments can inform semantic interpretation of document
- Techniques to separate documents into passages include
  - Rule-based systems based on clue words and phrases
  - Probabilistic techniques to separate fragments and to identify discourse segments (Oddy)
  - Lexical cohesion to identify fragments (TextTiling)

Texttiling

- Uses lexical cohesion to identify segments, assuming that each segment exhibits "lexical cohesion" within the segment, but is not cohesive across different segments
- Algorithm
  - Identifies candidate segments
  - Computes lexical cohesion score in each segment – Lexical cohesion score is the average semantic similarity of words within a segment
  - Identify boundaries by the difference of cohesion scores
  - NLTK has a text tiling algorithm available

Discourse Structure

- Human discourse often exhibits structures that are intended to indicate common experiences and respond to them
  - For example, research abstracts are intended to inform readers in the same community as the authors and who are engaged in similar work
  - Essay structure taught to high school students
  - Newspaper structure, where the story is given in several segments lending itself to shorter or longer versions

Discourse Relations

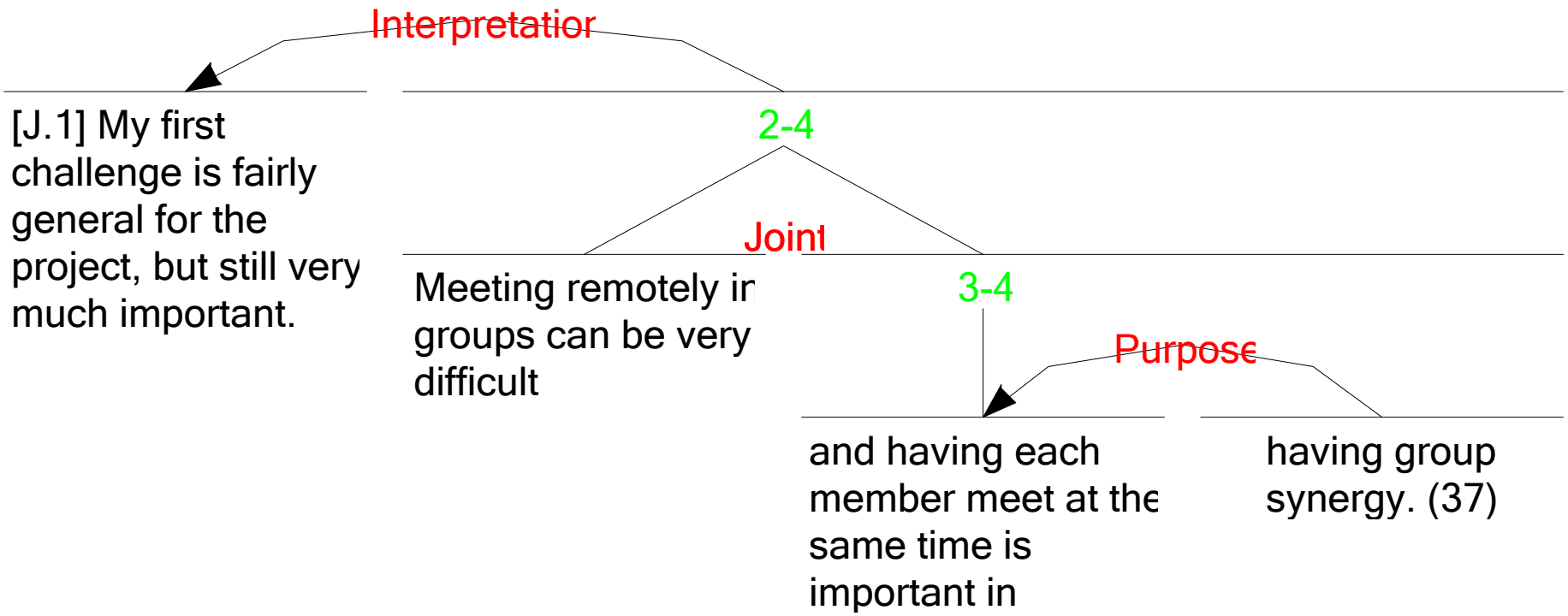
Rhetorical Structure Theory (RST): a theory of text organization created in the 1980s

Text units as *nuclear* and *satellites*

Three categories of relations:
subject matter relations,
presentational relations,
multinuclear relations

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

Rhetorical Structure Theory



RST Annotation In The Rationale Texts (Xiao & Conroy, 2017)

Rutgers Argument Mining Corpora
(Wacholder, Muresan, Ghosh, Aakhus, 2014)

Re: #18 You must be a shill for the RIAA and others. Truth is that file sharing in fact has been a boon to truly independent filmmakers and other non affiliated content producers, because they've become better known; it only really harms the big sharks, who have been robbing everyone blind and suppressing the little guy for decades.

Secondly, piracey is illegal. You aren't sharing anything. It's just a eufimisom for STEALING, and that's exactly what it is. It cheats the creator out of their money.

RST Annotation In The Rationale Texts (Xiao & Conroy, 2017)

Common RST relations in the callouts that have rationales

"Common" RST relations	Percentage in Coded RST relations
Justify	19.11%
Elaboration	13.31%
Conjunction	9.74%
Contrast	8.81%
Joint	7.73%
Evaluation	5.93%
Circumstance	4.75%
Condition	4.61%
Non Volitional Cause	3.81%
Concession	3.22%
	81.03%

Corpora	Percentage in coded RST relations
android	77.81%
ban	76.79%
ipad	81.82%
layoff	84.08%
twitter	82.05%

Computational Analysis: Discourse Parsing

- Discourse Segmentation
- Discourse Relation Detection
- Rhetorical Tree Building

Li, J., Li, R., & Hovy, E. H. (2014). Recursive Deep Models for Discourse Parsing. In *EMNLP* (pp. 2061-2069)

da Cunha, I. (2013). A Symbolic Corpus-based Approach to Detect and Solve the Ambiguity of Discourse Markers. *Research in Computing Science*, *70*, 95-106.

Bhatia, P., Ji, Y., & Eisenstein, J. (2015). Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

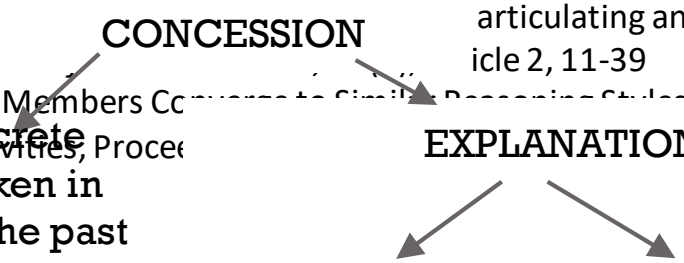
Rationale Detection: A Corpus-based Lexical Cue Graph Model (Khazaei & Xiao, 2015; Khazaei, Xiao, & Mercer, 2015)

Step 1: Identify the **RST relations that are commonly present** in the rationale texts

Three RST relations are reported to be commonly present in the rationales: **CIRCUMSTANCE, EVALUATION,** and **ELABORATION**

Step 2: Identify the **lexical cues for these RST relations** based on **two corpora**

1. **Xiao, L.,** & Carroll, articulating and sharing rationale: An empirical study. *International Jour* icle 2, 11-39
2. **Xiao, L.** (2013) Do Members Co[illegible] Reasoning Styles in Team[illegible] Study of Shared Rationale in Small Team Activities, Proce[illegible]



Energetic and concrete action has been taken in Colombia during the past 60 days against the mafiosi of the drug trade,

but it has not been sufficiently effective,

because, unfortunately, it came too late.

Corpora
- RST Discourse Treebank
- SFU Review Corpus

Explicit relations are the ones that are signaled by cues:
- ***lexical cues***
- punctuations
- mood
- ..

Corpora

- RST Corpus - News corpus
  - 385 Wall Street Journal Articles
- SFU Corpus – Reviews corpus
  - 400 reviews from movie, book, and products

Lexical Cue Extraction

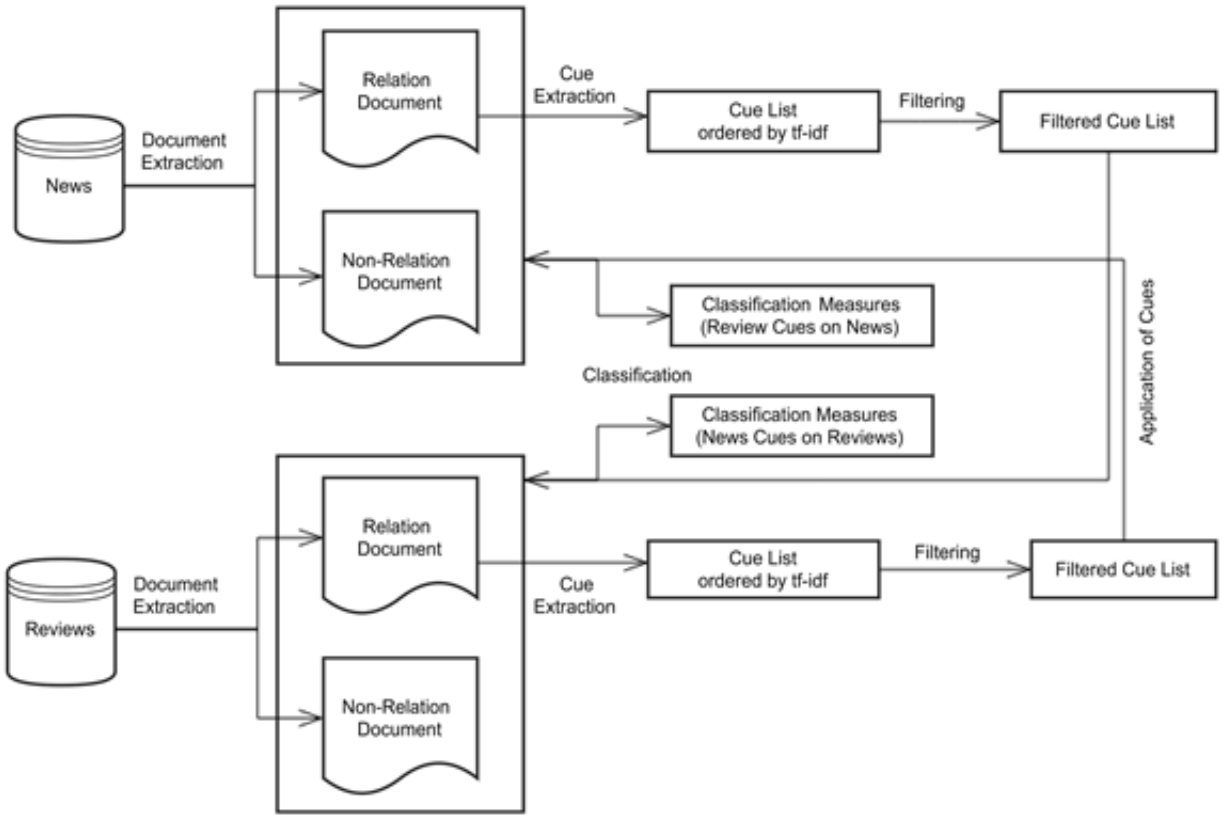
Biran And Rambow, 2011

- Extract all of the relation instances from the corpus and form a document named after the relation
- Extract all the n-grams from the document
- Calculate TF-IDF, sort the n-grams, and filter the list (to exclude any pronoun, or auxiliary verb)

Example:

Relation	**News**	**Reviews**
CIRCUMSTANCE	when, now, since	while, until, once
EVALUATION	good, high, well	nice, love it, impress
ELABORATION	who, which, as	which, where, as if

Experiment Description



Finding – CIRCUMSTANCE, ELABORATION, EVALUATION

- CIRCUMSTANCE
  - Heavily signaled
  - Signals are common discourse markers
  - Relatively genre-independent
- EVALUATION
  - Cues are not traditional markers
  - Dependent on the underlying genre, may or may not be signaled
- ELABORATION
  - Low weight and F-score
  - Corpus-based lexical cue approach is NOT recommended

Discourse Linguistics: Text Cohesion And Coherence

Cohesion and Coherence

- A text will exhibit unity / texture
  - on the surface level (cohesion)
  - at the meaning level (coherence)
- Halliday & Hasan's Cohesion in English (1976)
  - Categories of the surface level ties
  - Sets forth the linguistic devices that are available in the English language for creating this unity / texture
  - Identifies the features in a text that contribute to an intelligent comprehension of the text

Cohesive Relations

- Define dependencies between sentences in text.

"He said so."

"*He*" and "*so*" presuppose elements in the preceding text for their understanding

- This presupposition and the presence of information elsewhere in text to resolve this presupposition provide COHESION

- Part of the discourse-forming component of the linguistic system

- Provides the means whereby structurally unrelated elements are linked together

Six Types Of Cohesive Ties

- Grammatical
  - Reference
  - Substitution
  - Ellipsis
  - Conjunction
- Lexical
  - Reiteration
  - Collocation
- (In practice, there is overlap; some examples can show more than one type of cohesion.)

1. Reference

- items in a language which, rather than being interpreted in their own right, make reference to something else for their interpretation.

- Anaphora example with references (he, his, there) to previous text:

  "Doctor Foster went to Gloucester in a shower of rain. He stepped in a puddle right up to his middle and never went there again."

- Cataphora example with pronoun (he) referring to following text.

  When he visited the construction site last month, Mr. Jones talked with the union leaders about their safety concerns.

2. **Substitution**:

- a substituted item that serves the same structural function as the item for which it is substituted.

  Nominal – *one, ones, same*

  Verbal – *do*

  Clausal – *so, not*

  - *These biscuits are stale. Get some fresh ones.*
  - Person 1 – *I'll have two poached eggs on toast, please.*

    Person 2 – *I'll have the same.*

  - *The words did not come the same as they used to do. I don't know the meaning of half those long words, and what's more, don't believe you do either, said Alice.*

3. Ellipsis

- Very similar to substitution principles, embody same relation between parts of a text
- Something is left unsaid, but understood nonetheless, but a limited subset of these instances
  - *Smith was the first person to leave. I was the second* _____.
  - *Joan brought some carnations and Catherine* _____ *some sweet peas.*
  - *Who is responsible for sales in the Northeast? I believe Peter Martin is* _____.

4. Conjunction

- Different kind of cohesive relation in that it doesn't require us to understand some other part of the text to understand the meaning

- Rather, a specification of the way the text that follows is systematically connected to what has preceded

For the whole day he climbed up the steep mountainside, almost without stopping.

And in all this time he met no one.

Yet he was hardly aware of being tired.

So by night the valley was far below him.

Then, as dusk fell, he sat down to rest.

Now, 2 types of Lexical Cohesion

Lexical cohesion is concerned with cohesive effects achieved by selection of vocabulary

5. Reiteration continuum –

I attempted an ascent of the peak. _X__ was easy.

- same lexical item – *the ascent*
- synonym – *the climb*
- super-ordinate term – *the task*
- general noun – *the act*
- pronoun - *it*

6. Collocations

Lexical cohesion achieved through the association of semantically related lexical items

- Accounts for any pair of lexical items that exist in some lexico-semantic relationship, e. g.

- complementaries
  boy / girl
  stand-up / sit-down
- antonyms
  wet / dry
  crowded / deserted
- converses
  order / obey
  give / take
- pairs from ordered series
  Tuesday / Thursday
  sunrise / sunset
- part-whole
  brake / car
  lid / box
- co-hyponyms of same super-ordinate
  chair / table (furniture)
  walk / drive (go)

Uses Of Cohesion Theory

- Scoring text cohesiveness
  - Halliday & Hasan's theory has been captured in a coding scheme used to quantitatively measure the extent of cohesion in a text.
  - ETS has experimented with it as a metric in grading standardized test essays.

- Language generation and machine translation can use cohesion and coherence to build fluent texts

Building Semantic Representations

- When building a semantic representation of a text, the theory suggests how the system can recognize relations between entities.
  - Which entities in the text are related
  - How they are related
    - Particularly, coreference resolution finds all of the references to the "same" entity and groups them into clusters
- Information Extraction requires coreference resolution to build the relation triples

Lexical Chains

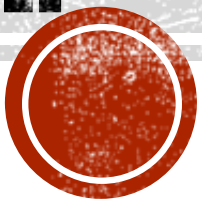
- Building lexical chains is one way to find the lexical cohesion structure of a text, both reiteration and collocation.
- A lexical chain is a sequence of semantically related words from the text
- Document can be viewed as a set of lexical chains
  - A kind of clustering of words based on semantic similarity
  - Each cluster can be viewed as a document "topic"
- Algorithm sketch:
  - Select a set of candidate words
  - For each candidate word, find an appropriate chain relying on a "relatedness" measure among members of chains
    - Usually semantic similarity between words
  - If it is found, insert the word into the chain.

Coherence Relations – Semantic Meaning Ties

- The set of possible relations between the meanings of different utterances in the text
- Hobbs (1979) suggests relations such as
  - **Result:** state in first sentence could cause the state in a second sentence
  - **Explanation:** the state in the second sentence could cause the first
    John hid Bill's car keys. He was drunk.
  - **Parallel:** The states asserted by two sentences are similar
    The Scarecrow wanted some brains. The Tin Woodsman wanted a heart.
  - **Elaboration:** Infer the same assertion from the two sentences.
- Textual Entailment
  - NLP task to discover the result and elaboration between two sentences
- The examination of transitional phrases and the measure of lexical overlap in our study (Khazaei, Xiao, & Mercer , 2017)

Khazaei, T., Xiao, L., & Mercer, R. (2017). Writing to Persuade: Analysis and Detection of Persuasive Discourse

Discourse Linguistics: Coreference Resolution



Anaphora/Reference Resolution

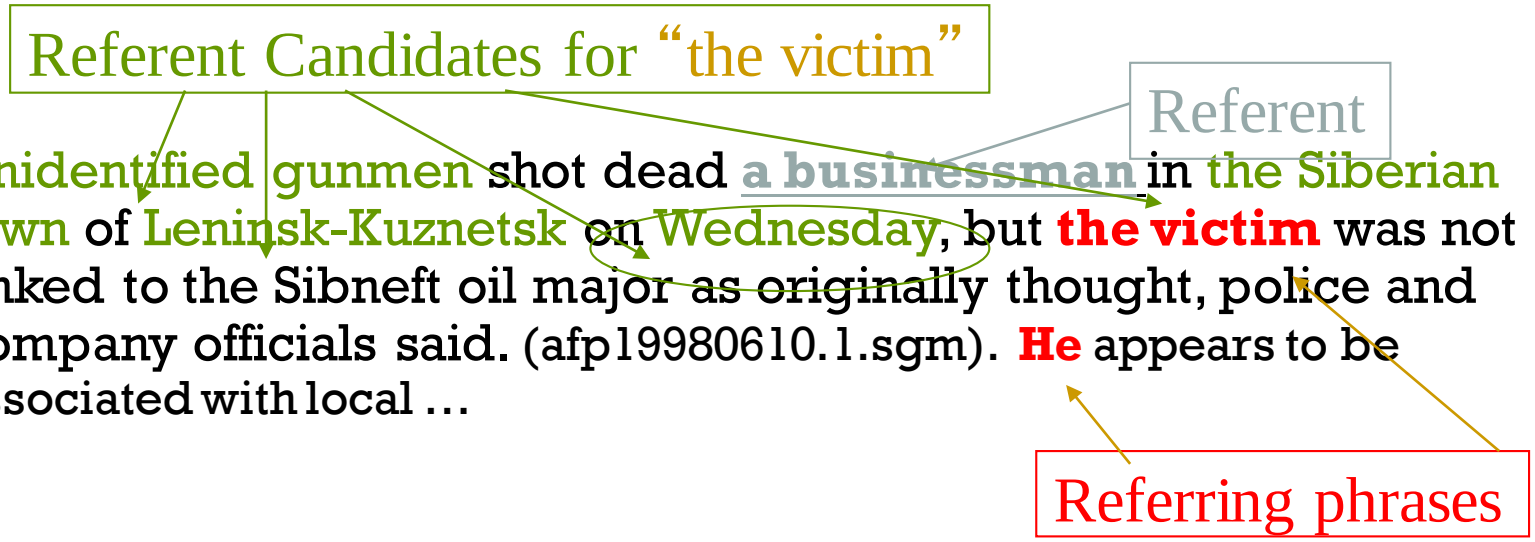
- A linguistic phenomenon of abbreviated subsequent reference
  - A cohesive tie of the grammatical and lexical types
    - Includes reference, substitution and reiteration
  - A technique for referring back to an entity which has been introduced with more fully descriptive phrasing earlier in the text
  - Refers to this same entity but with a lexically and semantically attenuated form

Types Of Entity Resolutions

- **Entity Resolution** is an ability of a system to recognize and unify variant references to a single entity.
  - Coreference algorithms usually performed within larger task of entity resolution
- 2 levels of resolution:
  - within document (includes **co-reference resolution**)
    - e.g. *Bin Ladin = he*
    - *his followers = they*
    - *terrorist attacks = they*
    - *the Federal Bureau of Investigation = FBI = F.B.I*
  - across document (or **named entity resolution**)
    - e.g. *Usama Bin Ladin =Osama Bin Ladin = Bin Ladin*
- **Event resolution** is also possible
  - Bejan, C. A., & Harabagiu, S. (2014). Unsupervised event coreference resolution. *Computational Linguistics*, *40*(2), 311-347.

Terminology Examples

- The referent for a referring phrase is found by the resolution algorithm among the candidates, previous noun phrases.



Reference Types

- An algorithm must first decide which are the referring phrases that must be resolved
  - Pronouns
  - Definite noun phrases (the)
  - Indefinite noun phrases (a, an)
  - Demonstratives (this, that)
  - Names
  - Others

Pronouns

- **Pronouns** refer to entities that were introduced fairly recently, 1-4-5-10(?) sentences back.
  - ***Nominative*** (he, she, it, they, etc.)
    - e.g. The German authorities said **a Colombian$_1$** who had lived for a long time in the Ukraine flew in from Kiev. **He$_1$** had 300 grams of plutonium 239 in his baggage.
  - ***Oblique*** (him, her, them, etc.)
    - e.g. Undercover investigators negotiated with **three members of a criminal group$_2$** and arrested **them$_2$** after receiving the first shipment.
  - ***Possessive*** (his, her, their, etc. + hers, theirs, etc.)
    - e.g. **He$_3$** had 300 grams of plutonium 239 in **his$_3$** baggage. The suspected smuggler$_3$* denied that the materials were **his$_3$**. ***(*chain)***
  - ***Reflexive*** (himself, themselves, etc.)
    - e.g. There appears to be a growing problem of **disaffected loners$_4$** who cut **themselves$_4$** off from all groups .

Definite Noun Phrases – The X

- Definite reference is used to refer to an entity identifiable by the reader because it is either
  - a) already mentioned previously (in discourse), or
  - b) contained in the reader's set of beliefs about the world (pragmatics), or
  - c) the object itself is unique. (Jurafsky & Martin, 2000)
- E.g.
  - Mr. Torres and his companion claimed **a hardshelled black vinyl suitcase$_1$**. The police rushed **the suitcase$_1$** (a) to **the Trans-Uranium Institute$_2$** (c) where experts cut **it$_1$** open because they did not have the combination to the locks.
  - **The German authorities$_3$** (b) said **a Colombian$_4$** who had lived for a long time in **the Ukraine$_5$** (c) flew in from Kiev. He had **300 grams of plutonium 239$_6$** in his baggage. **The suspected smuggler$_4$** (a) denied that **the materials$_6$** (a) were his.

Indefinite Noun Phrases – A X, Or An X

- Typically, an indefinite noun phrase introduces a new entity into the discourse and would not be used as a referring phrase to something else
  - The exception is in the case of cataphora:
    A Soviet pop star was killed at a concert in Moscow last night. Igor Talkov was shot through the heart as he walked on stage.

Unidentified gunmen shot dead a businessman in the Siberian town of Leninsk-Kuznetsk on Wednesday, but **the victim** was not linked to the Sibneft oil major as originally thought, police and company officials said. (afp19980610.1.sgm). **He** appears to be associated with local …

Demonstratives – This And That

- Demonstrative pronouns can either appear alone or as determiners

  this ingredient, that spice

- These NP phrases with determiners are ambiguous
  - They can be indefinite

    I saw this beautiful car today.
  - Or they can be definite

    I just bought a copy of Thoreau's Walden. I had bought one five years ago. That one had been very tattered; this one was in much better condition.

Names

- Names can occur in many forms, sometimes called name variants.
  Victoria Chen, Chief Financial Officer of Megabucks Banking Corp. since 2004, saw her pay jump 20% as the 37-year-old also became the Denver-based financial-services company's president. Megabucks expanded recently . . . MBC . . .
  - (Victoria Chen, Chief Financial Officer, her, the 37-year-old, the Denver-based financial-services company's president)
  - (Megabucks Banking Corp. , the Denver-based financial-services company, Megabucks, MBC)
- Groups of a referrent with its referring phrases are called a coreference group.

Unusual Cases

- Compound phrases
  John and Mary got engaged. They make a cute couple.
  John and Mary went home. She was tired.
- Singular nouns with a plural meaning
  The focus group met for several hours. They were very intent.
- Part/whole relationships
  John bought a new car. A door was dented.

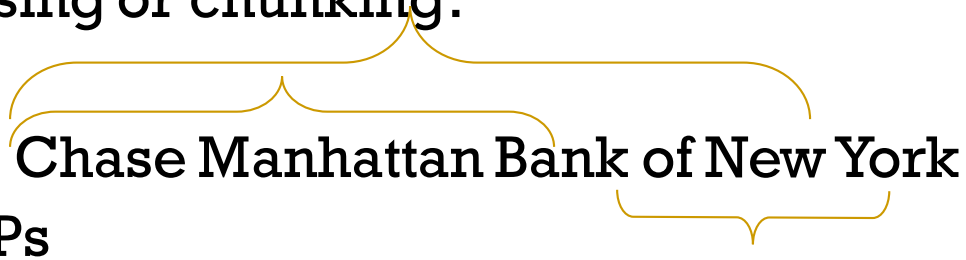
  Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer.

Approach To Coreference Resolution

- Naively identify all referring phrases for resolution:
  - all Pronouns
  - all definite NPs
  - all Proper Nouns
- Filter things that look referential but, in fact, are not
  - e.g. geographic names, *the United States*
  - pleonastic "it", e.g. *it's 3:45 p.m., it was cold*
  - non-referential "it", "they", "there"
    - e.g. *it was essential, important, is understood,*
    - *they say,*
    - *there seems to be a mistake*

Identify Referent Candidates

- All noun phrases (both indef. and def.) are considered potential referent candidates.
- A referring phrase can also be a referent for a subsequent referring phrases,
  - Example: (omitted sentence with name of suspect) He had 300 grams of plutonium 239 in his baggage. The suspected smuggler denied that the materials were his. (chain of 4 referring phrases)
- All potential candidates are collected in a table collecting feature info on each candidate.
- Requires either parsing or chunking:
  - chunking
    - e.g. the Chase Manhattan Bank of New York
  - Note nesting of NPs



Some Features

- Define features between a referring phrase and each candidate
  - Number agreement: plural, singular or neutral
    - He, she, it, etc. are singular, while we, us, they, them, etc. are plural and should match with singular or plural nouns, respectively
    - Exceptions: some plural or group nouns can be referred to by either it or they
      IBM announced a new product. They have been working on it …
  - Gender agreement:
    - Generally animate objects are referred to by either male pronouns (he, his) or female pronouns (she, hers)
    - Inanimate objects take neutral (it) gender
  - Person agreement:
    - First and second person pronouns are "I" and "you"
    - Third person pronouns must be used with nouns

Some Features (cont'd)

- Binding constraints
  - Reflexive pronouns (himself, themselves) have constraints on which nouns in the same sentence can be referred to:
    John bought himself a new Ford. (John = himself)
- Recency
  - Entities situated closer to the referring phrase tend to be more salient than those further away
    - And pronouns can't go more than a few sentences away
- Grammatical role, sometimes approximated by Hobbs distance
  - Entities in a subject position are more likely than in the object position

Approaches

- Train a classifier over an annotated corpus to identify which candidates and referring phrases are in the same coreference group
  - Evaluation results (for example, Vincent Ng at ACL 2005) are on the order of F-measure of 70, with generally higher precision than recall

- Lassalle, E., & Denis, P. (2015, January). Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures. In *AAAI* (pp. 2274-2280)
  - employ latent trees to represent the full coreference and anaphoricity structure of a document

- Prokofyev, R., Tonon, A., Luggen, M., Vouilloz, L., Difallah, D. E., & Cudré-Mauroux, P. (2015, October). SANAPHOR: ontology-based coreference resolution. In *International Semantic Web Conference* (pp. 458-473). Springer International Publishing
  - Leverages Semantic Web technologies

Summary Of Discourse Level Tasks

- Dialogue structure (discourse segmentations, discourse relations, text coherence)
- Document structure
  - Recognizing known structure, for example, abstracts
  - Separating documents according to known structure
- Named entity resolution across documents