Text to Networks and Semantic Networks

Prof. Kathleen M. Carley

kathleen.carley@cs.cmu.edu

From Texts to Networks
From AutoMap/NetMapper to ORA

• Extract a network from texts using a text mining tool
  - E.g. AutoMap or NetMapper
  - Easy approach is D2M – data-to-model

• Load relational data into ORA
  - Visualize the network
  - Generate specialized reports
    • Hot topics report
    • Semantic network report
  - Generate other reports
Network Text Analysis: two-step, iterative process

1. Text pre-processing:
   - Condense data to terms relevant for domain and research question
     - Normalization
     - Reduction
   - Supports the task of finding meaningful interpretations of texts
     Natural Language Processing and Information Extraction preconditions for NTA

2. Link Formation:
   - Determine how the relevant concepts will be linked into ties

Data To Model Procedure

- Preprocessors
  - Clean and compartmentalize texts
  - Speeds processing
- D2M script
  - Common bi-grams
  - Thesauri
  - Deletes
  - Relation extraction
  - Ontology cross-classification
- Postprocessing
  - Attribute addition
  - Geo-location
  - Membership and belief inference
  - Event inference (still in development)
Key: Coding Choices

• Text pre-processing:
  – Condense data to concepts relevant for domain and research question
  – Supports the task of finding meaningful interpretations of texts

• Statement Formation:
  – Determine how the relevant concepts will be linked into statements
  – Analysts: tailor them to research question

Text pre-processing: Filters

• Deletion
  – remove non-content bearing words such as determiners, conjunctions, articles, prepositions
  – Highly efficient

• Thesauri
  – Convert diverse concepts to the same concept e.g., aliases
  – Enables theme construction
Text pre-processing: 
Apply Background Information

- Thesauri
  - A thesaurus converts from one concept to another
  - Column N: relevant text-level concepts in the
  - Column N+1: corresponding higher-level concepts
    - Represents text-level concepts in a more abstract or general way
    - A single higher level concept typically has multiple entries in the thesaurus.

- Translation
  - Translations convert from other languages to English

- Attributes
  - Known attributes of the concept
  - E.g. for people this might include first name, gender, role
  - E.g. for locations this might include country

Universal and Domain

Terminology in information fields might depend on data and research question that is highly specialized

- Universal
  - Thesauri – General conversions that most people want
  - Delete List – General stop words, numbers, measures and temporal indicators

- Domain
  - Thesauri – Special conversions for the specific data and research question not covered by universal or that alter the universal choice
  - Delete List – Special concepts that are uninteresting in the domain
    - E.g. all concepts with low TFIDF or low frequency or isolates

- Thesauri’s over-ride delete lists
- Domain thesauri over-rides universal thesauri
- Universal delete list over-rides domain delete list
Support Tools for Creating Domain Files

- Thesaurus creation support:
  - Union Concept List
  - N-grams (especially bigrams)
  - Feature Selection (e.g. tf-idf)
  - External sources (CIA World Fact Book, WordNet)
  - Bootstrapping

- Translation
  - People
  - Online word lists
  - Google translate
  - Other online translators

- Attributes
  - Coding schemes
  - Online structured data

Pre-processing

- “John Doe is actively involved in several industry and civic associations.”
- Deletion:
  - Remove non-content bearing concepts
  - High data reduction rates
  - Delete List: is, in, and
  - Delete List creation support: concept list per text and corpus (Zipf’s law)
  - Adjacency:
    - Direct: John Doe actively involved several industry civic associations.
    - Rhetorical: John Doe xxx actively involved xxx several industry xxx civic associations.
Pre-processing

- **Stemming**
  - Detect inflections and derivations of concepts
  - Convert each concept into related morpheme
  - John Doe is actively involved in several industry and civic associations.
  - Porter: John Doe be activ involv in sever industri and civic associ.
  - Krovetz: John Doe is actively involve in several industry an civic association.


- **Thesauri**
  - Convert concepts with more abstract concepts (generalization thesaurus) or meta-network entities (meta-network thesaurus)
  - John Doe is actively involved in several industry and civic associations.
    - Generalization thesaurus
    - meta-network thesaurus
    - John Doe/ John_Doe
    - John_Doe/ agent
    - involved/ engage
    - engage/ task
    - Industry/ economy
    - economy/ resource
    - civic associations/ community
    - community/ organization
Getting Started - Data Collection

- Newspaper articles
- Books
- Journal articles
- Abstracts
- Interviews
- Essay Questions
- Email
- Web pages
- Dictionary entries
- Scripts
- Concept circles
- Power point presentations

Any type of raw text

Representation Scheme

<table>
<thead>
<tr>
<th>DNA Term</th>
<th>Term</th>
<th>Graphic</th>
<th>Equivalent Ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node</td>
<td>Link</td>
<td>Concept</td>
<td>Word Phrase</td>
</tr>
<tr>
<td>Link</td>
<td>Link</td>
<td>Link</td>
<td>Sentence</td>
</tr>
<tr>
<td>Dyad</td>
<td>Statement</td>
<td>Statement</td>
<td>Implication</td>
</tr>
<tr>
<td>Network</td>
<td>Map</td>
<td>Focus</td>
<td>Relation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Idea</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fact</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Definition</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mental Model</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Frame</td>
</tr>
</tbody>
</table>
Concept Circle

**Idea**

List of concepts
Place concepts in circle
Have respondent draw lines between linked concepts

**When Appropriate**

Want to cue respondent
Want only predefined concepts
Less than 50 concepts

**Usage Hint**

Train respondents on small task (10 concepts)
Go over map with respondent

---

Concept Circle - Example

**Directions:** These words have been mentioned in class lectures over the past semester. Please draw a line between pairs of words which you believe should be connected. It is important that all connections that you intend to make be clear and easy to see. Please do not draw so many lines on any one worksheet that you cannot easily see how you've connected those words.
When Respondent Draws Lines

- Place strength on lines
- Place arrows on lines for causality
- Place marker on lines for type of link

Application Process

- Can be applied by interviewer during interview
- Can be done as reading text

Variations:

Link formation

- In order to extract semantic network from raw or pre-processed data we need to link the nodes
- Multiple approaches exists:
  - Distance based approach: Windowing:
    - Text Unit
    - Window Size
    - Adjacency

Syntactic Based Link Formation

- Distance based approach, Windowing:
  - Text Unit (text, paragraph, sentence)
  - Window Size (2 to N)
  - Adjacency (direct or rhetorical)
- Thesaurus:
  - leader/leader;
  - involved/engage;
  - industry/industry;
  - civic associations/civic_association
- Coding Choices:
  - Sentence, thesaurus content only, rhetorical adjacency, window size sentence
- TEXT: Leader xxx actively involved xxx several industry xxx civic associations.

From Texts to Networks

- Distance based approach, Windowing:
  - Text Unit (text, paragraph, sentence)
  - Window Size (2 to N)
  - Adjacency (direct or rhetorical)

<table>
<thead>
<tr>
<th>Leader</th>
<th>xxx</th>
<th>actively</th>
<th>involved</th>
<th>xxx</th>
<th>several</th>
<th>industry</th>
<th>xxx</th>
<th>civic associations</th>
</tr>
</thead>
</table>

Relational data:
- no thesaurus.
- Leader - actively actively - involved Involved - several

Thesaurus content only, direct adj. WS 3 or rhet. adj. WS 10:
Extracted Network as Mental Models

- Semantic Network (Map)
  - Representation of a sample of what people have in their minds and use to make sense of their surroundings.
  - Cognitive constructs that reflect the subjects' knowledge and information about a certain topic.


One-mode networks: semantic networks

1. One mode networks: semantic networks, cognitive/mental models:
   - Represent reality that people have in their minds and use to make sense of their surroundings.
   - Cognitive constructs that reflect the subjects’ knowledge and information about a certain topic.

- Applications:
  - Interviews: individual and group-level
  - Self-presentations: annual reports, self-applications, mission statements
  - Groups: (Shared) Situational Awareness

Hints

- If using exact words
  - use a frequency level – e.g., only if used more than "3" times, or used by 10% people
- If generalizing
  - create thesaurus file with examples of how to code words, be very general
- General
  - code 2 or 3 texts first to make sure you like set up
- Either code for specific concepts or use delete lists
  - Delete non content bearing words
  - Delete off topic words
- Comment
  - there will be a limited number of words per topic/community – usually 2-3000
  - Some limit to most frequent – this can be misleading
    - May contain off topic
    - Rank may be very different if you use thesauri as key concepts generally have more alternative forms than do other words
Illustrative Applications

- Enron
  - Diesner Carley
- Al Qaeda
  - Carley
- Decision Making
  - Carley
- Learning to Write
  - Palmquist PhD
- Risk Perception
  - Bostrom PhD
- World Bank & Poverty
  - Saburi PhD
- Authorship
  - Lewis
- Impact of Email on Invisible College
  - Gardner & Carley
- Robotics in Science Fiction
  - Student Project
- What Children Remember
  - Cicourel
- COVID-19
  - This institute

Text pre-processing: Negative Filters

- Deletion
  - remove non-content bearing words such as determiners, conjunctions, articles, prepositions
  - Highly efficient
- Adjacency:
  - Direct: drop delete words and move word right and left to delete word directly next to each other
  - Rhetorical: insert a placeholder where a delete word was removed in order to maintain original distance of all words
Impact of Coding Choices on Network

Impact of statement formation settings on number of unique statements

Semantic Network as Meta-Networks

- We might also want to know about other nodes
  - Who is linked to what organizations
  - Who has access to what resources
  - Who possesses what knowledge
  - Who is involved in what tasks
  - Who was at what event
  - What resources are needed for what tasks
  - ...

- We might also want to know about other attributes
  - Who has what personal characteristics
  - What types of locations exist
  - ...
Semantic Network as Meta-Networks

- Implement meta-network model as ontology: Meta-Network Text Analysis
  - Extract structure of social and organizational systems from corpora
  - Provide relational data that can be further analyzed with other tools and methods
- Use your own ontology
  - Rename existing categories


Multi-mode Networks (aka Meta-Networks)

- Social Networks
  - Supports asking traditional/baseline questions:
    - Who is connected to whom?
    - What type of relationship?
  - With meta-network data we can move beyond traditional questions and also ask:
    - Which agent or group is located where, has access to what resources, possesses what knowledge, is involved in what tasks, has what personal characteristics, ...?
- AutoMap:
  - Use meta-network as is
  - Use self-modified meta-network
  - Use your own ontology
- NetMapper
  - Use meta-network as is
  - Use category 1 to specify your ontology then switch in ORA
  - Relabel categories in ORA

Extraction of Relational Data From Texts – Toy Example

Example from UN News Service (New York), 12-28-2004:

"Jan Pronk, the Special Representative of Secretary-General Kofi Annan to Sudan, today called for the immediate return of the vehicles to World Food Programme (WFP) and NGOs."

Extraction of Relational Data From Texts – Two Approaches

- Any type of network:
  - **Locate** / **identify** “relevant” nodes, e.g. knowledge
  - Sparse

- Ontologically coded, socio-technical networks:
  - **Classify** relevant nodes according to some ontology or taxonomy

Default Ontology

- **Named Entities**
  - **Agent**
    - Individual actors
    - Specific – unique often with first and last name - Jamie O’Connor
    - Generic – non-unique and often a role - haberdasher
  - **Organization**
    - Groups, corporations, populations
    - Specific – unique - IBM
    - Generic – a type - Non-government organization
  - **Location**
    - Places things can be at
    - Specific – unique with lat and lon or place on map – United States of America
    - Generic – may be at multiple locations – hill
  - **Event specific, generic**
    - Major happenings that impact groups
    - Specific – occur once – World War I
    - Generic – multiple occurrences - Tornado

- **General Entities**
  - **Knowledge**
    - Branches of knowledge
    - Topics of interest
  - **Resources**
    - Things that are not purely mental – disease, food, wire
  - **Tasks**
    - Activities – eat
  - **Beliefs**
    - "isms" - Catholicism
    - Sentiment – positive, negative
    - Belief statements – right to bear arms
Ontological Text Coding

- Requires mapping of words to categories (*agent*, *event*, *knowledge*, *location*, *organization*, *resource*, *task*, *attribute*)

1. Explicit mapping (positive filters)
   - Deterministic, incomplete, outdated, spelling variations, typos
   - Exploit available data (CIA world fact book, lists, ...)

2. Regular expressions
   - Identifying relevant strings of text (particular characters, words, or patterns) (JGSoft: RegexBuddy)
   - KWIC (keyword in context) ([A-Za-z0-9] = Alphanumeric chars.)
   - URL’s, dates, phone numbers


Thesauri Format

```
<table>
<thead>
<tr>
<th>conceptFrom</th>
<th>conceptTo</th>
<th>Ontology</th>
<th>nodeType</th>
<th>Category</th>
<th>First_Name</th>
<th>Last_Name</th>
<th>Gender</th>
<th>Hon</th>
<th>Verif</th>
<th>Acr</th>
</tr>
</thead>
<tbody>
<tr>
<td>mine_clearance</td>
<td>mine_clearance</td>
<td>task</td>
<td>belief</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parents</td>
<td>parent</td>
<td>agent</td>
<td>generic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pat_Mcginn</td>
<td>Pat_Mcginn</td>
<td>agent</td>
<td>specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

- Parents: Pat Mcginn
- Gender: Male
- Country: United States of America
- Verification Level: 1
Window Based Location of Links

- John_Doe is actively engage in several economy and community.

```
<table>
<thead>
<tr>
<th>engage</th>
<th>economy</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Doe</td>
<td>community</td>
</tr>
</tbody>
</table>
```

Window Based Location of Links with Ontology

- agent is actively task in several resource and organization.

```
<table>
<thead>
<tr>
<th>task</th>
<th>resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>agent</td>
<td>organization</td>
</tr>
</tbody>
</table>
```
Example Text Files

Hisham Al Hussein
... the Philippine government booted the second secretary at Iraq's Manila embassy, Hisham Al Hussein, on February 13, 2003, after discovering that the same mobile phone that reached his number on October 3, 2002, six days later rang another cell phone strapped to a bomb at the San Roque Elementary School in Zamboanga.

Abu Madja and Hamsiraji Ali
That mobile phone also registered calls to Abu Madja and Hamsiraji Ali, leaders of Abu Sayyaf, Al Qaeda's Philippine branch.

Abdurajak Janjalani
It was launched in the late 1980s by the late Abdurajak Janjalani, with the help of Jamal Mohammad Khalifa, Osama bin Laden's brother-in-law.

Hamsiraji Ali
... Hamsiraji Ali, an Abu Sayyaf commander on the southern island of Basilan, bragged that his group received almost $20,000 annually from Iraqis close to Saddam Hussein.

Place Concepts in Meta-Network

<table>
<thead>
<tr>
<th>Agents</th>
<th>Knowledge</th>
<th>Resources</th>
<th>Tasks</th>
<th>Locations</th>
<th>Organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>abu_madja</td>
<td>school</td>
<td>bomb</td>
<td>bomb</td>
<td>basilian</td>
<td>al_qaeda</td>
</tr>
<tr>
<td>bin_laden</td>
<td>phone</td>
<td></td>
<td>philippine</td>
<td></td>
<td>bu_sayyaf</td>
</tr>
<tr>
<td>hamsiraji_ali</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>manila</td>
</tr>
<tr>
<td>hisham_hussein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>zaboanga</td>
</tr>
<tr>
<td>janjalani</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jami_khalifa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>saddam_hussein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
... and knowledge and resources

Semantic Networks

Prof. Kathleen M. Carley

kathleen.carley@cs.cmu.edu
Semantic Network

- Network of concepts
- Sometimes referred to as
  - a mental model
  - An information network
- Generally extracted from texts
- Variations on nodes
  - Words –vs- concepts
  - topics –vs- concepts
- Features of links
  - Co-use or logic or semantic relations
  - Strength – generally represents frequency
    - In text or across texts

Illustrative Semantic Network

- Healthcare
- User needs
- User community
- Sale techniques
- Gyms
- Computer capabilities
- New products
- Programming
- Product problems
- Inventory Systems
- Accounting Systems
- Product information
Why compare texts as networks: Locating Common Ground

- Do people use same words
- Do people use same words in same way
- Do people evoke the same concepts
- Do people link the same concepts in the same way
- Are the same disciplines employed
- Are the same journals, trade magazines, etc. read

Information Network: Text & Graphics

Student A: I found that scientists engage in research in order to make discoveries and generate new ideas. Such research by scientists is hard work and often involves collaboration with other scientists which leads to discoveries which make the scientists famous. Such collaboration may be informal, such as when they share new ideas over lunch, or formal, such as when they are co-authors of a paper.

Student B: It was hard work to research famous scientists engaged in collaboration and I made many informal discoveries. My research showed that scientists engaged in collaboration with other scientists are co-authors of at least one paper containing their new ideas. Some scientists make formal discoveries and have new ideas.
Network Approach to Communicative Power

- Three key communicative dimensions
  - Intensity/consensus – weighted degree or number of graphs of individual’s mental models or number of reports the tie is present in
  - Conductivity – indegree*outdegree or betweenness
  - Density – degree
- Most words will be low on all three dimensions
- Words that stand out on any one dimension will have extra communicative power
- Messages containing such words will be more persuasive

Distributional Information

- Communicative power is a function of reach
- Reach depends on the network position of author and of concept
- Every concept has a position in the meta-network
  - Document by concept – link is number of times document uses concept
  - Concept by concept – link is number of documents that use both concepts (i.e., the author’s/document’s semantic network)
- Concepts vary in their meta-network position
- Network position of concept determines its communicative power
Cube analysis of communicative power: Concept Embedding

The Cube

- Consensus: Usage across documents
- K-Betweenness Semantic Network
- Total Degree Semantic Network
- Standard symbols
- Emblems
- Factoids
- Stereotypes
- Buzzwords
- Placeholders
- Ordinary words
- Allusions
- Low
- High
General: Ordinary Word

- sense of frustration
- expansion
- maddeningly
- contradictory evidence

Factoid

- Nov. 22, 1963
- 4 Density
- 3 Conductivity
- 4 Consensus
**Emblem**

- Kevin Costner
- fights injustice
- good guy
- actor
- hero
- decency
- 5 Density
- 25 Conductivity
- 5 Consensus

**Stereotype**

- Capra-style Hero
- 16 Density
- 0 Conductivity
- 10 Consensus
**Standard Symbol**

- President
- 15 Total Degree
- 120 Betweenness
- 8 Frequency

**Allusion**

- Grassy Knoll
- 15 Total Degree
- 0 Betweenness
- 8 Frequency
Cube

**Purpose**
Generate network measures on concept position

**Output**
Concept by value matrix

**Have Ready**
- A semantic network
- A meta-network that is the union across documents
  - Document x concept
  - Concept x concept
- Often we treat document as location
- Concept are type knowledge

**Illustrative Concept - Car**

Car

3 Promotes Social Interaction
3 Encourage hall activities
5 Resource

5 Total Degree
6 Betweenness
1 Frequency
Illustrative Concept — Gnerd

- self-punishing, works too hard
- personality style of speech
- friendly, gets along
- interacts with students
-acher
- shares personal, expansive, stories
- adaptable
- liberal, hip
- sensitive, perceptive
- accessible

Illustrative Concept — Hacker

- proud to be a third easter
- jokes around, expressive, dramatic
- liberal, hip
- sensitive, perceptive
- accessible

Density: 32
Conductivity: 207
Betweenness: 0.969
Frequency: 0.969
Distribution of 3E concepts

3e Cube

Count

0
50
100
150

general
dateoids
stereotypes
placeholderssymbols
Highly Communicative Concepts

Degree Betweenness Consensus

```
"car"   5.000  6.000  1.000  factoids
"kids"  5.000  6.000  1.000  factoids
"intelligence"  26.000  120.000  0.846  prototypes
"interesting"  28.000  132.000  0.643  prototypes
"phantom"  48.000  176.000  0.938  stereotypes
"hacker"  32.000  207.000  0.969  stereotypes
"participates"  32.000  220.000  0.969  stereotypes
"teaching-abi"  36.000  224.000  0.889  stereotypes
"hall-member"  32.000  231.000  0.969  stereotypes
"expert"  31.000  238.000  0.968  stereotypes
"lookupableto"  51.000  270.000  0.882  symbols
"gnerd"  40.000  336.000  0.950  symbols
"accessible"  41.000  378.000  0.925  symbols
"gets-along"  64.000  735.000  0.781  placeholders
"fits-in"  67.000  780.000  0.821  placeholders
```

Cube and Social Knowledge

```
fits-in  (placeholder)
(67,780,0.821)

gets-along (placeholder)
(64,735,0.781)

insists on quiet (factoid)
(10,24,1.0)

mellow (factoid)
(19,70,0.947)

tolerant (ordinary)
(23,76,0.870)

egocentric (ordinary)
(10,16,0.800)

number students ok TC (factoid)
(14,33,0.929)
```
Summary of Cube analysis

- Texts can be coded as networks
- Networks represent the mental model
- Coding texts as mental models focuses on meaning
- Concepts have position in texts
- Examining positions focuses on communicative power
- Classifying concepts creates Meta-Networks