 Semantic Networks: Introduction

- Structured representations of information and knowledge
- Originally meant to be used for reasoning and inference
  - E.g. theory of spreading activation

Collins and Loftus (1975)
Semantic Networks: Key Ingredients and Representation

1. Syntax: specifies eligible types of nodes and edges
   - Nodes: represent concepts
   - Edges: typed or associative
   - Optional: meta-data

2. Specification of meaning/semantics of nodes, links, graph

3. Inference rules

Using Semantic Networks: Extract Meaning

- **Prediction**: forecast the set of concepts that will be evoked when a certain note is activated
- **Disambiguation**: network clustering to identify different aspect of the meaning of a concept.
- **Summarization**: retrieve a concept’s ego-network to distill the essence of some data in concise and structured form.
Semantic Networks: State of the Art and Construction

- Woods 1975: Attractive notion, but lack theoretical grounding and rigor in representational conventions
- Today: anywhere between strict and highly flexible notion
- Translation or transformation of input data
  - Translation:
    - Goal: convert natural language input into isomorphic, structured representations
    - Does not scale well
  - Transformations:
    - Abstraction process that preserves and reveals entities and relations that are explicitly or implicitly represented in input data.
    - Goal: reduce dimensionality of input data to capture relevant structural relations and provide input for inference
- Can be constructed from text data, among other sources

From Words to Networks: Association Networks


Association Networks: 
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
- Typical data sources:
  - Interviews
  - Self-presentations (annual reports, applications, mission statements)
- Methods for data collection:
  - Interviews, questionnaires paired with text analysis
  - Pile sort


From Words to Networks: 
Association Networks: Relation Extraction

Language Networks: Introduction

- **Assumption:** Language, information and knowledge can be modeled as relational data
- **Fact:** Collection and storage of large volumes of text data cheap, easy and efficient
  - Interviews, books, news wire articles, legal documents, annual reports, data from web 1.0 (web sites) and web 2.0 (emails, blogs, chats, ...)
- **Need:** Methods and tools for automated, robust and reliable knowledge discovery and reasoning about information, incl. network structures, from text data.
- **Challenge:** Effective, efficient and controlled extraction of relevant (user-defined) instances of categories (node and edge) from unstructured, natural language text data.

Language Networks: Originate from many Disciplines

- Respective theories and methods developed across many disciplines:
  - Artificial Intelligence (e.g. Sowa)
  - Cognition and Linguistics (e.g. Collins)
  - Communications (e.g. Doerfel, Monge)
  - Political Science (e.g. Schrodth)
  - Sociology (e.g. Carley, Mohr)
  - Computer Science (e.g. McCallum)
Basic Types of Information in Text Data

- **Morphology**: structure of words  
  - E.g. spelling, inflections, derivations
- **Syntax**: relationships between words  
  - E.g. parts of speech tagging
- **Semantics**: meaning of language  
  - E.g. word sense disambiguation, grammars
- **Pragmatics**: language in context and social use of language  
  - E.g. sentiment analysis, discourse analysis
- **Relation Extraction (talk and tutorial)**: borrows from all of the above

Text Data Applications

- **Data analysis**
- **Network Data Construction & Analysis**
  - Scalable, reliable, robust methods & technologies
  - Network data
  - Data analysis
- **Applications**
  - Answer substantive questions about networks
  - Fill databases
  - Forecasting: Explore future behavior and what-if scenarios
  - Input to other computations, e.g. machine learning

From Words to Networks, Jana Diesner, 2013
From Words to Networks: Example: Sudan

**Task:** Develop, evaluate and apply method and technology for extracting socio-technical network data from large-scale text corpora to answer questions about the Sudan.

Semantic Networks: Knowledge Representation

- Declarative or definitional semantic networks
- Key ingredient: Ontologies

Porphyry of Tyre (234–305 A.D.)
What node classes to consider?  
**Ontology**

- **Who?** (people, groups)
- **Where?** (places)
- **Why?** (beliefs, sentiments, mental models)
- **What?** (tasks, events)
- **How?** (resources, knowledge)
- **When?** (time)

How to find and categorize nodes in text data?  
**Basic recipe for probabilistic solution**

- Get some labeled ground-truth data (BBN)
- Build a classifier/model (h) that for every sequence of words (x) and label per word (y) predicts one category per word (y = h(x)), incl. for new and unseen text data
- Exploit clues from text data (lexical, syntactic, statistical)
- Train and validate the model
- Get good accuracy (compare to intercoder reliability) (we made model available in end-user product AutoMap)
- Apply prediction model to text data (~ 80,000 files)
- Link nodes (e.g. based on co-occurrence, proximity)
- Network data! Analysis!


How to find and categorize nodes in text data?

- Sequences of $x$ (words) and $y$ (label)
  $P(x,y)$: generative models, e.g. Hidden Markov Model (HMM).
- $P(y|x)$: conditional models, e.g. Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF).

- CRF:
  - Consider arbitrarily large bag of features
  - Consider and any property of $x$, incl. long-range features

Model relationship among hidden states ($y$) as Markov Random Field (MRF) conditioned on observed data ($x$) (Lafferty et al. 2001)

Compute conditional distribution of entity sequence $y$ and observed sequence $x$ as normalized product of potential functions $M_i$:

$$M_i(y_{i-1}, y_i | x) = \exp\left(\sum_a \lambda_a f_a(y_{i-1}, y_i, x) + \sum_b \mu_b g_b(y_i, x)\right)$$

$$P(y|x) = \frac{\prod_{i=1}^{n-1} M_i(y_{i-1}, y_i | x)}{\prod_{i=1}^{n-1} M_i(x)_{\text{start,stop}}}$$

- Edge and transition features plus node and emission features
- $f, g$: boolean feature vectors with learned weights
- Tool: CRF project page, training data: BBN

From Words to Networks, Jana Diesner, 2013
### How good is it?

<table>
<thead>
<tr>
<th>Class model</th>
<th>Class</th>
<th>Specificity</th>
<th>Subtype</th>
<th>Example</th>
<th>States, Edges</th>
<th>Time (300)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td>agent</td>
<td>11, 121</td>
<td>17.5 hours</td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td>x</td>
<td></td>
<td>agent, specific</td>
<td>16, 256</td>
<td>1.25 days</td>
</tr>
<tr>
<td>3</td>
<td>x</td>
<td>x</td>
<td></td>
<td>agent, political</td>
<td>32, 1024</td>
<td>3.1 days</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>x</td>
<td></td>
<td>agent, spec., pol.</td>
<td>45, 2025</td>
<td>5 days</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Accuracy</th>
<th>Training Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>large</td>
<td>small</td>
</tr>
<tr>
<td>Syntax features (POS)</td>
<td>small</td>
<td>small</td>
</tr>
<tr>
<td>Lexical features (dict, hard match)</td>
<td>large</td>
<td>small</td>
</tr>
<tr>
<td>Iteration rate</td>
<td>large</td>
<td>large</td>
</tr>
<tr>
<td>Complexity of class model</td>
<td>small</td>
<td>large</td>
</tr>
</tbody>
</table>

### Network Data! Analysis!

**Activity:**

- Degree Centrality
- Betweenness Centrality
- Eigenvector Centrality

**Control:**

- President North: Known performer
- President South: Now established
- Legacy of religious leaders
- Presence of neighboring presidents

**Close to power:**

- Darfur conflict: Continuous civil war (since 1993)
- Comprehensive Peace Agreement: Garang 1st VP, followed by Kiir Autonomus South Sudan
- SPLA withdraws from government
- Votum in South Sudan about Separation

---

From Words to Networks, Jana Diesner, 2013
• Strong presence of armed forces
• Strong influence of external groups
• Within top 10 Sudanese groups:
  – Dinka, Nuer (ethnic groups/tribes)

From Words to Networks, Jana Diesner, 2013
Sudan: Results:
Conflict, War and Resources

- Conflict: Agriculture, Livestock (farmers vs. herders)
- War: Land Resource (concept of *dar*)
- Conflict and War: Oil, Civic, Transportation

Sudan: Results (Key themes)

<table>
<thead>
<tr>
<th>Topics with highest Activity</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conflict</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cultural</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Peace Making</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Biomes/Land Cover</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conflict</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Kinship</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cultural</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Peace Making</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bridging topics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ideology Religion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Welfare</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Security Forces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Political</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Water Management</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hmm, Relation Extraction looks like a nice idea. How accurate are your results?

I fine-tuned our method and technology based on the F-values and feedback from SMEs.

But the F only shows the increase in accuracy over a baseline or benchmark. Maybe we need to ask a different question…

The F values tell me all I need to know.

Problem Statement and Question

• **Problem**: Impact of methodological choices about relation extraction on network data and analysis results mainly unknown

• **Question**: How do network data and analysis results differ depending on methodological choices?

• **Who cares?**
  – Increased comparability, generalizability, transparency of methods and tools
  – Increased control and power for developers and users
  – Supports drawing of reasonable and valid conclusions

Diesner J (2013) From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data. Künstliche Intelligenz/ Artificial Intelligence. DOI: 10.1007/s13218-012-0225-0
# Methods for Going from Words to Networks

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference(s)</th>
<th>Automation</th>
<th>Abstraction</th>
<th>Generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mental Models (Spreading Activation)</td>
<td>(Collins &amp; Loftus 1975)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Discourse Representation Theory</td>
<td>(Kamp 1981)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Centering Resonance Analysis</td>
<td>(Corman et al. 2002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Mind maps</td>
<td>(Buzan 1974)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Concept maps</td>
<td>(Novak &amp; Gowin 1984)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Hypertext</td>
<td>(Trigg &amp; Weiser 1986)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Qualitative text coding (Grounded Theory)</td>
<td>(Glaser &amp; Strauss 1967)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Definitional semantic networks incl. text coding with ontologies</td>
<td>(Fellbaum 1998)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Frames</td>
<td>(Minsky 1974)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Network Text Analysis in social science</td>
<td>(Carley &amp; Palmquist 1991)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Event Coding in pol. science</td>
<td>(King &amp; Lowe 2003, Schrodt et al. 2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Probabilistic graphical models</td>
<td>(Howard 1989, Pearl 1988)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparing Methods: Data

<table>
<thead>
<tr>
<th></th>
<th>Sudan Corpus</th>
<th>Funding Corpus</th>
<th>Enron Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genre</strong></td>
<td>Newswire</td>
<td>Scientific Writing</td>
<td>Emails</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>80,000 articles</td>
<td>56,000 proposals</td>
<td>53,000 emails</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>LexisNexis</td>
<td>Cordis</td>
<td>FERC/SEC</td>
</tr>
<tr>
<td><strong>Time span</strong></td>
<td>8 years</td>
<td>22 years</td>
<td>4 years</td>
</tr>
<tr>
<td><strong>Text-based networks</strong></td>
<td>Article bodies</td>
<td>Project description</td>
<td>Email bodies</td>
</tr>
<tr>
<td><strong>Meta-data network</strong></td>
<td>Index terms (knowledge)</td>
<td>Index terms (knowledge) and collaborators (social)</td>
<td>Email headers (social)</td>
</tr>
</tbody>
</table>

- All: large scale, over time, open source data from different domains
Comparing Methods: Automated node prediction in application domains

- Method: systematic evaluation of auto-generated thesauri on all 3 datasets
- No meaningful differences in accuracy across domains, time, writing styles
  - Technology generalizes AND generalizes better than manually built thesauri
  - Creation and refinement more efficient (time) and effective (finding nodes) than manually built thesauri
- Subtype “specific” more unique/different instances, but “generic” far more total instances
  - Rethink focus of network analysis:
    - More references to roles and collectives than to individuals
    - Importance of extracting unnamed entities
    - “Specific” instances lower accuracy than “generic” ones due to data sparsity

How do relation extraction methods compare?

- Ground truth data (SME) hardly resembled by analyzing text bodies, not at all by meta-data networks
  - SME in TextM: 53% nodes 20% links
  - SME in TextA: 11% nodes, 5% edges
- Agreement in structure and key entities mainly function of:
  - Size of extracted graph
  - External material/ sources used
  - Post-processing/ cleaning
    - Agreement can be coincidental if no proper word sense disambiguation performed
    - Type of network: semantic versus typed
Methods Assessment

<table>
<thead>
<tr>
<th>Type</th>
<th>Text-Based Networks</th>
<th>Meta-Data Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social networks</td>
<td>- Substantial overlap TextM and TextA, esp. key players (identity, rank)</td>
<td>- Small overlap in key entities with text-based networks</td>
</tr>
<tr>
<td></td>
<td>- Localized view on geopolitical entities and culture</td>
<td>- Key players: major international agents, hardly localized views</td>
</tr>
<tr>
<td>Semantic/knowledge networks</td>
<td>- Minimal overlap between manual and automated node identification</td>
<td>- Seem more informative (crafted mini-summaries)</td>
</tr>
<tr>
<td></td>
<td>- Gist of information in terms of common sense, highly salient entities</td>
<td>- Less coreference resolution issues</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Minimal overlap with text-based</td>
</tr>
</tbody>
</table>

For more complete view, combine automated text-based with meta-data network
Cover common/highly salient terms and entities and domain-specific ones

References

Acknowledgements

• Sudan study: This work was supported by the National Science Foundation (NSF) IGERT 9972762, the Army Research Institute (ARI) W91WAW07C0063, the Army Research Laboratory (ARL/CTA) DAAD19-01-2-0009, the Air Force Office of Scientific Research (AFOSR) MURI FA9550-05-1-0388, the Office of Naval Research (ONR) MURI N00014-08-11186, and a Siebel Scholarship. Additional support was provided by CASOS, the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, ARI, ARL, AFOSR, ONR, or the United States Government.

• I thank Nikita Basov and the Centre for German and European Studies at the St. Petersburg State University, Russia, for hosting this workshop.

Thank you!

Q&A

• For questions, comments, feedback, follow-up – now and later:
  Jana Diesner
  jdiesner@illinois.edu
  Phone: (412) 519 7576
  Web: http://people.lis.illinois.edu/~jdiesner