Relation Extraction from Texts and Computational Integration of Words and Networks

Carnegie Mellon

institute for SOFTWARE RESEARCH

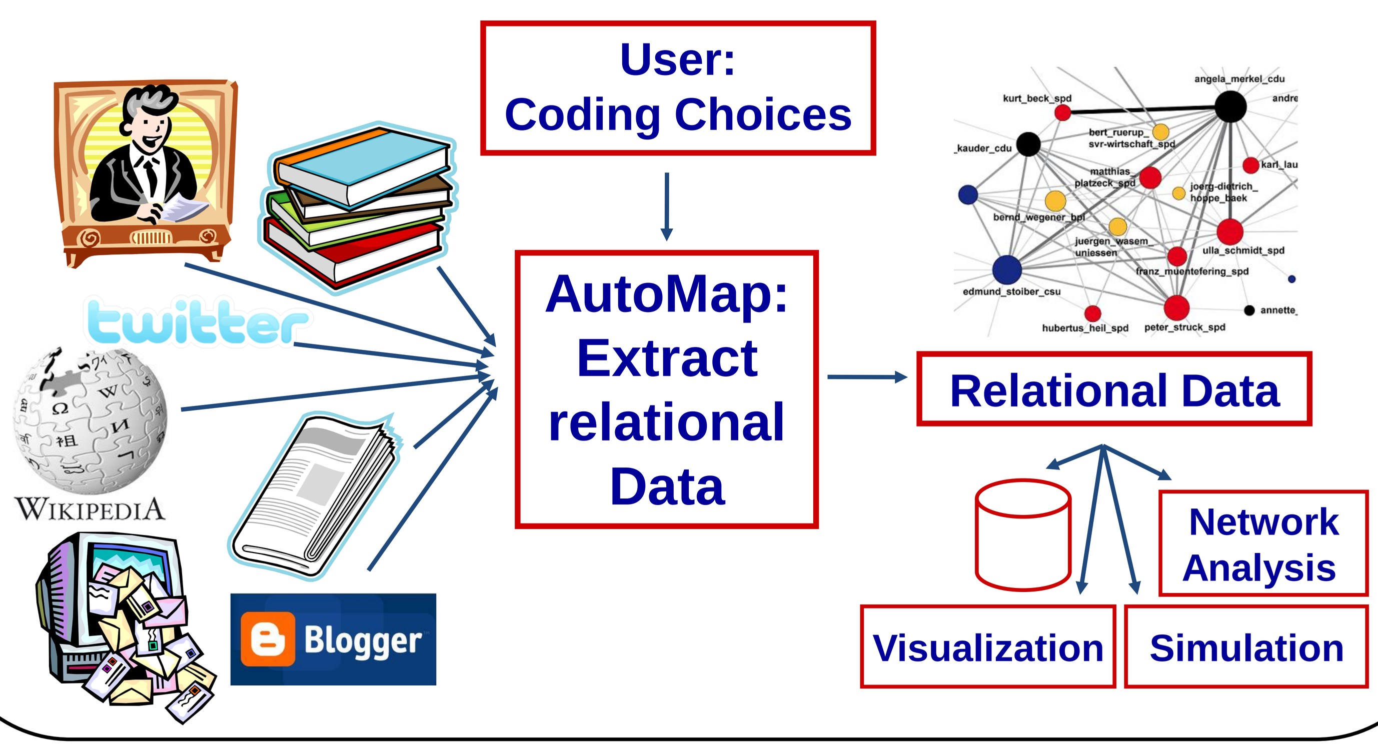
Ph.D. Program in Computation, Organizations & Society

Jana Diesner

diesner@cs.cmu.edu

Prof. Kathleen M. Carley

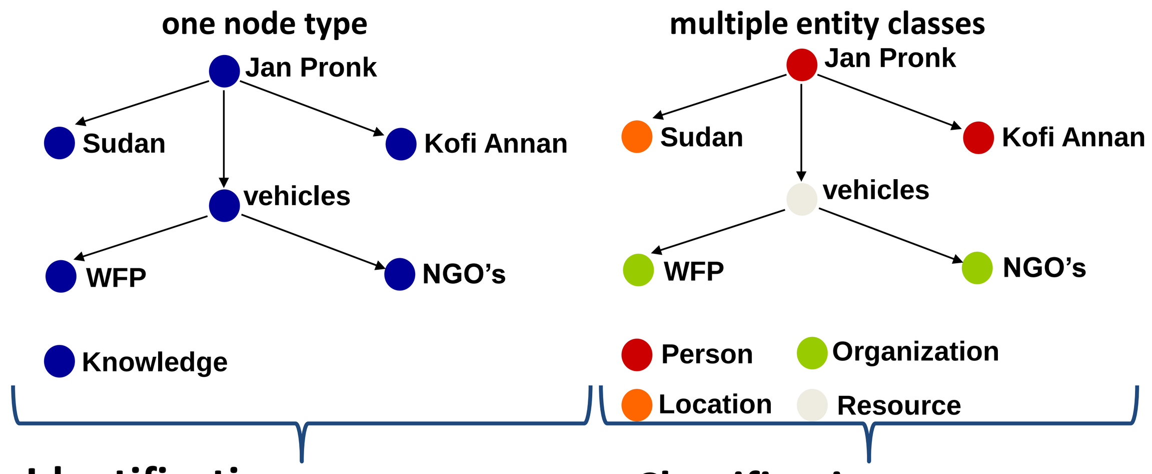
kathleen.carley@cs.cmu.edu



Illustrative Toy Example:

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

proximity-based extraction of relational data :

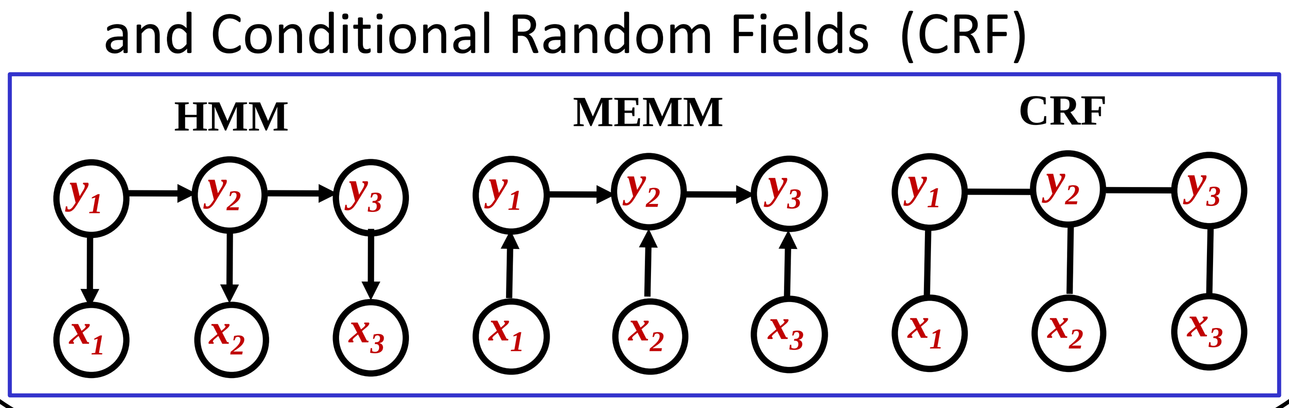


Identification: For relational data with at least one node type: **Locate/ identify** relevant nodes (may be multi-word units)

Classification: For ontologically coded networks: **Classify** relevant nodes according to an ontology or taxonomy

Development of Computational Solutions

- Utilize techniques from Machine Learning and Artificial Intelligence
- Deploy and develop supervised and semi-supervised **sequential stochastic learning techniques** in order to train classifiers and build models that generalize to new data
- Construct a classifier *h* that for every sequence of ***(x, y)*** (joint probability) (where *x* = words per sequence and *y* = corresponding category) or ***(x|y)*** (conditional probability) predicts a sequence ***y = h (x)*** for any sequence of *x*, incl. new and unseen data
- We work with Generative (aka discriminative) models: *P(x,y),* such as Hidden Markov Model (HMM), and Conditional models: *P(y|x)*, such as Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF)



References:

Carley, K. M., Diesner, J., Reminga, J., & Tsvetovat, M. (2007). Toward an interoperable dynamic network analysis toolkit. *Decision Support Systems (DSS),* 43(4), 1324-1347.

Diesner, J., & Carley, K. M. (2009). He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis. Proceedings of *IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA)*, Ottawa, Canada, July 2009.

Diesner, J., & Carley, K. M. (2005). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. In V. K. Narayanan & D. J. Armstrong (Eds.), *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations* (pp. 81-108). Harrisburg, PA: Idea Group Publishing.

Diesner, J., & Carley, K. M. (2008). Conditional Random Fields for Entity Extraction and Ontological Text Coding. *Journal of Computational and Mathematical Organization Theory (CMOT), 14*, 248 - 262.

Diesner, J., Carley, K. M., & Katzmair, H. (2007). The morphology of a breakdown. How the semantics and mechanics of communication networks from an organization in crisis relate. XXVII Sunbelt Social Networ Conference, Corfu, Greece, May 2007.

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory (CMOT), 11*(3), 201-228.

McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. *ACM Queue, 3*(9), 48-57.

Natural Language Processing and Relational Extraction Routines in AutoMap

- **Stemming**: Convert words into their morphemes.
- **Reduction and Normalization**:
  - Negative filters such as delete lists and removal of symbols
  - Positive filters such as spelling correction and assigning synonyms to unique key concept
- **Part of Speech Tagging**: Assign a single best word class to every word.
- **Anaphora Resolution**: Convert personal pronouns into entity or entities that a pronoun refer to.
- **Feature Identification:** Automatically find the most important terms in a dataset.
- **Named Entity Extraction**: Identify relevant types of information that are referred to by a name, such as people, organizations, and locations.
- **Ontological Text Coding**: Identify and classify instances of pre- or user-defined node classes, such as Named Entities, resources, tasks, and time.
- Identification of and reasoning about **node and edge attributes**, such as demographic data, beliefs, and types of relationships.
- **Email Data Analysis**: Extract and combine different types of networks, such as social networks and knowledge networks, from emails.
- **Entropy Assessment**: Determine the variability of a text or text set with respect to its vocabulary.
- Classical **Content Analysis**.
- Read and write data and processing material from and to a default or user-specified **database**.

Example: Conditional Random Fields for Entity Extraction

- Identify and classify words that represent instances of entity classes of models or ontologies that **deviate** from classical set of Named Entities.
- Crucial step for coding texts as social-technical networks according to domain-specific ontologies and for advanced modeling of complex and dynamic real-world organizations or networks.
- Model relationship among y_i and y_{i-1} as Markov Random Field conditioned on *x*
- Conditional distribution of entity sequence *y* given observation sequence *x* computed as normalized product of potential functions M_i:

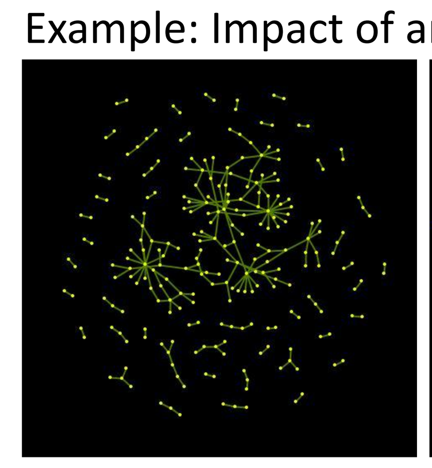
$$M_i(y_{i-1}, y \mid x) = \left(\exp\left(\sum_\alpha \lambda_\alpha f_\alpha(y_{i-1}, y_i, x) + \sum_\beta \mu_\beta g_\beta(y_i, x)\right)\right)$$

$$p_\theta(y \mid x) = \frac{\prod_{i=1}^{n+1} M_i(y_{i-1}, y_i \mid x)}{\prod_{i=1}^{n+1} M_i(x)_{start,stop}}$$

- Conditional probability of label sequence *P(y|x)*, where both *x* and *y* are arbitrarily long vectors (consider arbitrarily large bag of features (> 10,000))and any property of *x*, such as long-distance information)

Evaluation

- Rigorous assessment of the impact of information and relation extraction techniques on relational data and respective interpretations of socio-technical networks
- Example: Impact of anaphora and coreference resolution:



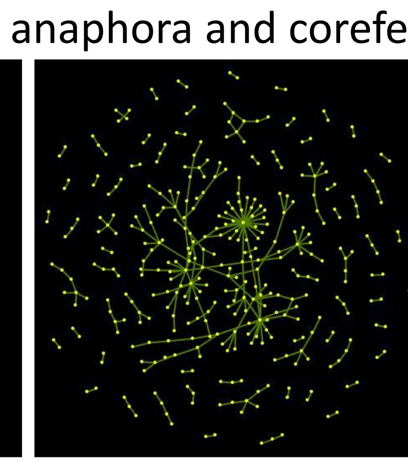


Table: Impact of AR, CR on edge level

Routine	measurement	newswire	newspaper	broadcast
raw	unique nodes	4715	4884	3743
	total node weight	5774	5916	4536
AR	unique nodes	4599	4682	3659
	node weight reduction rate	2.5%	4.1%	2.2%
CR	unique nodes	3324	3213	2835
	node weight reduction rate	29.5%	34.2%	24.3%
AR+CR	unique nodes	3050	2894	2596
	node weight reduction rate	35.3%	40.7%	30.6%
	node weight reduction rate from AR to AR+CR	5.8%	6.5%	6.4%

Visualization of relations in broadcast data of the ACE2 corpus (NIST, LDC): raw data (left image) and after applying anaphora and coreference resolution (right image), showing links with strength > 1, isolates are hidden

This work is part of the Dynamics Networks project at the Center for Computational Analysis of Social and Organizational Systems (CASOS) of the School of Computer Science (SCS) at Carnegie Mellon University (CMU). Support was provided, in part, by National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship (IGERT) program in CASOS (DGE-9972762). Additional support was provided by the Army Research Institute (W91WAW07C0063), the ONR (N00014-06-1-0104), and the ONR MURI (N000140811186). The views and proposal contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied by the Army Research Institute, of the Office of Naval Research, the National Science Foundation, or the U.S. government.

www.casos.cs.cmu.edu

National Science Foundation WHERE DISCOVERIES BEGIN

CASOS Summer Institute 2010