Social Network Extraction from Texts: Thesis Proposal

Apoorv Agarwal
Department of Computer Science, Columbia University

Overall Goal

Extract a social network from text where nodes are people and links are *social events*

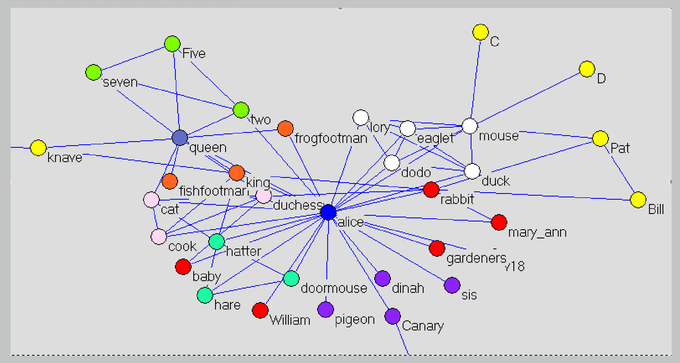


Figure: Social structure of *Alice in Wonderland*

Impact

- Social network analysis community
- Study of literary and journalistic texts
- Linguistics (add frames to FrameNet)

Social Events (Agarwal et. al. 2010)

Social Event: An event between two people or group of people where at least one party is aware of the other party and aware of the event. Types:

- Interaction event (INR): both parties mutually aware
- Observation event (OBS): only one party aware of the other

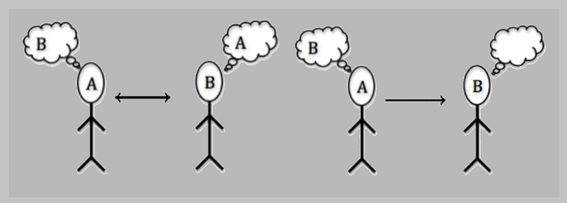
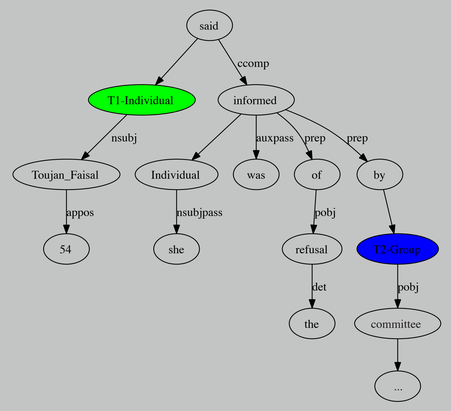


Figure: Interaction (INR) and Observation (OBS) social events respectively

- Subtypes of INR: INR.Verbal.Far, INR.Verbal.Near, INR.NonVerbal.Far, INR.NonVerbal.Near
- Subtypes of OBS: Physical proximity (PPR), Perception (PCR - *watch on T.V.*), Cognition (COG - *thinking*)

Social Event Extraction (Agarwal and Rambow, 2010)



- Tree structures:
  - ▷ Phrase Structure Tree (PET)
  - ▷ Dependency Word Tree (DW)
  - ▷ Dependency Grammatical Relation Tree (GR)
  - ▷ Dependency Grammatical Relation Word Tree (GRW)
- Sequence structures:
  - ▷ SK1: *T1-Individual Toujan_Faisal 54 said Individual she was informed of the refusal by an T2-Group Interior Ministry committee*
  - ▷ SqGRW: *Toujan_Faisal nsubj T1-Individual said ccomp informed prep by T2-Group pobj committee*
- Kernels:
  - ▷ Subset Tree (SST): used for PET
  - ▷ Partial Tree (PT): used for all other structures

Except SqGRW, all the above structures and PT are due to work by Nguyen and Moschitti, 2009

Future Work 1: Domain adaptation

- Current training data: English part of Automatic Content Extraction (ACE) 2005 multilingual news corpus
- Need domain adaptation to port the technique to richer sources of social networks like literary and historical texts
- Notable work by Hal Daumé III, 2007: present a straightforward kernelized version of domain adaptation

$$K'(T_1, T_2) = \begin{cases} 2K(T_1, T_2) & \text{if same domain} \\ K(T_1, T_2) & \text{if different domain} \end{cases}$$

Future Work 2: Scaling convolution kernels

$$f(x) = \sum_{i=1}^{N_s} \alpha_i y_i K(s_i, x) + b \quad \text{(Burges 1998)} \quad (1)$$

$$K(T_1, T_2) = \sum_{s} h_s(T_1) h_s(T_2) \quad \text{(Collins and Duffy 2002)} \quad (2)$$

$$f(x) = \sum_{i=1}^{N_s} \alpha_i y_i \sum_{s} h_s(s_i) h_s(x) \quad \text{(combining 1 \& 2)}$$

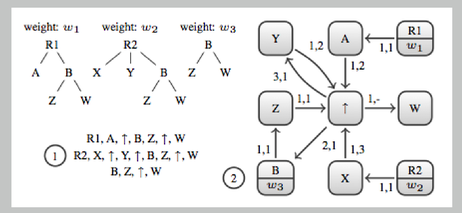
$$f(x) = \sum_{s} \sum_{i=1}^{N_s} \alpha_i y_i h_s(s_i) h_s(x) \quad \text{(interchanging summations)}$$

Questions:

- Can we restrict the set s to limited number of structures?
- Can we distribute the calculation of $f(x)$ over machines?
- Can we stop the calculation of $f(x)$ early?

Future Work 3: Interpreting convolution kernels

Linearization of SST kernel (Pighin and Moschitti, 2009):



(Figure taken from Pighin and Moschitti, 2009)

Output: a set of fragments that were given high weight in the high dimensional kernel space
Problem: thousands of fragments - interpretability is hard
Proposal: Define meaningful syntactic and semantic abstractions and attempt to categorize the set of fragments into these abstract classes

Future Work 4: Applications

- Predict hierarchy of Enron email corpus: capture "who talks to whom about whom" type of links
- Use phrase-level polarity analysis (Agarwal et. al 2009, 2011) to add polarity on edges of the network

http://www.cs.columbia.edu/∼apoorv apoorv@cs.columbia.edu,