Comparison of Data Development Tools for Populating Cognitive Models in Social Simulations

McKaughan, Daniel
Comparison of Data Development Tools for Populating Cognitive Models in Social Simulations

LCDR Daniel McKaughan
Agenda

- Introduction
- The Problem
- The Tools
- The Methods
- The Present
- The Finish
Introduction

• Who is speaking to you?
The Problem: Populating the Models

- Social Sim Goal: Represent groups of people (societies) as complex adaptive systems (CAS)
  - a complex, self-similar collection of interacting adaptive agents
- Desire accurate representation and analysis of society’s beliefs, values, and interests (BVI’s)
- Made up of actors which represent individuals making up population
- Actor’s must be informed of issues relevant within its group (How do we do this?)
Data & Scenario Development—CG Model

**Collect Data**
- Sources: DoD, academia, open press, polling data, experts, etc.
- Data: narratives, ASCOPE, PMESII-PT, SWEAT-MUS, observable behaviors.

**Analyze Data**
- Identify demographic dimensions & population groups; develop narrative identities; derive belief, values, or interest, & issue stances.
- Define stereotypes; define Bayesian belief networks and starting conditions; define agent behaviors; define agent count, disposition & location.
- Identify essential goods & services, providers, location & capacity of infrastructure.
- Identify and define social networks.
- Identify threat groups & other actors that influence the population; derive their motivations, beliefs, history, techniques, capabilities, & goals.

**Scenario Development**
- Implement of the civilian population, other actors, essential goods & services, and the supporting infrastructure.
- Develop scripted categories of events or semi-autonomous behaviors to support actors courses of action; social network interactions between population groups & actors; case files that impact beliefs & behaviors based on events or population behaviors.

**Iterative process focused on defining the population to obtain an accurate representation of relevant BVI and assessing how actions from and interactions of other actors within the conflict ecosystem impact related BVI, issue stances, and behaviors.**
The Tools

• Survey Data Development Tool
  – A Java based program that reads surveys as .csv files
  – Allows for selection of target question or response
  – Ranks terms either by Chi-Square comparison (Categorical) or Variance (Continuous)
  – Produces case files

• Sandia’s Text ANaLysis Extensible librarY
  – Gather documents using web crawlers or other means
  – Builds textual profile for each set of pre-categorized documents
  – Matches and scores new documents from corpus against all profiles using log entropy and cosine similarity functions
  – Marks document categorizations based on thresholds set by user
The Tools-Survey Data Development Tool

The Tools-Survey Data Development Tool is a software application designed to facilitate the development of survey data. The tool provides a user-friendly interface for creating and customizing survey data, allowing users to select and manipulate various data types and settings. The tool is particularly useful for researchers and data analysts who need to generate and manipulate survey data for analysis or other purposes.
Survey Data Case File Generator tool develops case files that represent opinion samples for each stereotype from survey data. Given a target term(s) or expected distributions for the term(s), the tool ranks the term(s) by comparing chi-square values (categorical terms) or variance (continuous terms) and develops case files. Set of case files using selected term(s) that represent opinion samples for each stereotype.
The Tools-STANLEY

Categorization scores from previous step

How well document matches the example corpus as a whole

Thresholds for matching to be set by user

How well document matches variable 1 and 2 example documents

How well document matches all variable example documents
**Agent Initialization Through Text Analysis**

*Purpose: To develop, implement, and verify a modular pipeline that uses text analysis for social simulation model initialization.*

- Primary Performer: Sandia National Laboratories.
- Agent initialization through text analysis process:

![Diagram of the process]

**Corpus Collection**
- Cognitive Crawler

**Document Categorization**
- Document Categorizer
  - A
  - B
  - C

**Agent Initialization**
- Observation Generator

**Requirements:**
- Create modular pipeline for agent initialization process.
- Provide initial test implementations for each module.
- Run and measure pipeline implementation using a test case.
The Methods

• Survey Data Development Tool
  – Gathered survey data and found applicable question to model
    • “Are you currently satisfied or dissatisfied with the state of the nation today”
    • Pew Research Council surveys conducted Jan-Aug 2010
  – Prepared survey data
    • Distinguished demographics for use in stereotype generation
    • Made software changes to tool’s code
  – Ran tool to develop case files
  – Developed scenario for use in CG model

– where I am now…
– Run CG model
– Compare results
The Methods

- **STANLEY**
  - Determined suitable problem
    - Can STANLEY aid in determination of correlation between Mid-East turmoil and oil prices in U.S. during early 2011?
  - Web crawl to gather documents about turmoil and documents about oil prices
    - *where I am now…*
  - Develop corpus of documents to test against
  - Run STANLEY
  - Score the documents
  - Determine any correlation
The Present

- Lots of time changing and modifying the Survey tool
  - Did not take strings
  - Spaces
  - Naming convention
- Finishing up input file
- Preparing CG

- Still running the crawlers for STANLEY
  - Massive amounts of results
    - 70,000 links from Forbes.com alone
  - Attempting to time stamp the articles to track oil prices
    - Match changes in Middle East to increasing oil prices
The Future

...
Questions