A Journey from Absolute Zero: The HLTCOE KELVIN System

6 November 2012

JHU: Paul McNamee, Ves Stoyanov, James Mayfield
UMBC: Tim Finin, Tim Oates
UMD: Tan Xu, Douglas Oard
Loyola: Dawn Lawrie

human language technology
center of excellence

It was a dark, cold, and stormy end of June

- Cast: 7 PhDs and 1 graduate student
  - Median age ~45 years
  - No single programming language in common
  - Lots of folks very confident in their programming skills, but "those who can't do, teach..."
- Distractions
  - TAC Entity Linking (English / Spanish)
  - TREC KBA
  - JHU SCALE 2012
  - ACM SIGIR
  - Misc: other projects, vacation schedules, adorable children

We were in desperate need of a strategy... And a system...

human language technology
center of excellence

Talk Outline

- **KELVIN: Knowledge Extraction, Linking, Validation, and INference**

- **Keys to Success**
- **Pipeline**
  - **Various Components Explained**
- **Limited Analysis**
- **Experimental Conditions**
- **Live Demo #1: Query Engine**
- **Live Demo #2: Browser**

Keys to Success

- Major Keys
  - Theft
  - Picnics
  - Reptiles
  - Intimacy
  - Safety Net

- Minor Keys
  - Glue
  - Lack of Shame
  - Tools
  - Practice

Theft (and repurposing)

- Relation Extraction
  - BBN SERIF
  - BBN FACETS
  - CUNY Slot Filling Toolkit
- Cold Start Validator
  - Kindly coerced into producing ‘inverses’
- HLTCOE CALE Entity Linker
  - Added to baseline runs
- Stanford NLP TIMEX normalizer

human language technology
center of excellence

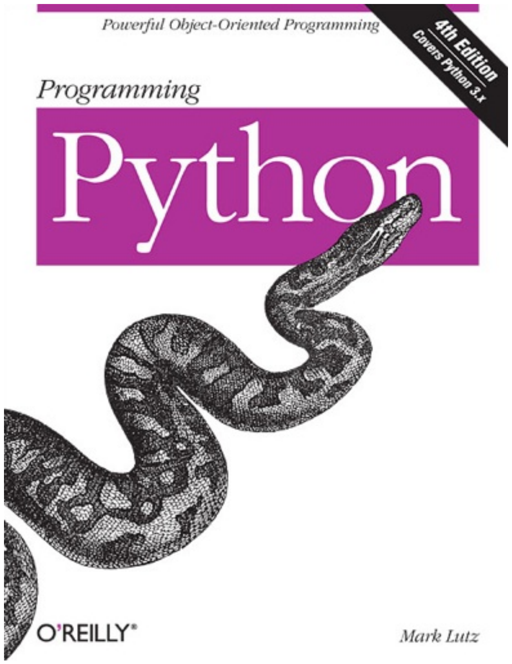
Picnics

- What works for graduate students appears to work for professors: offer food
  - Held ~9 "hack fests" from 10am to 5pm where 4+ members of our team worked around a conference table on laptops
  - Lunch provided
- Then spent a few days apart, working individually, and repeat process
- Surprisingly productive

human language technology
center of excellence

Reptiles

- **Python**
  - **6017 lines**
- **Perl**
  - **2665 lines (88% is Cold Start validator)**
- **Java**
  - **234 lines**

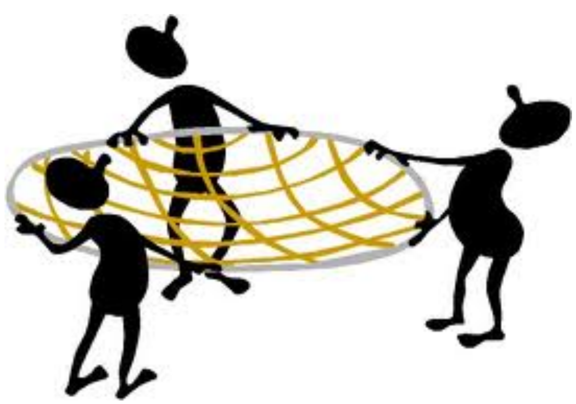


Intimacy

- It was useful having a task organizer working in the building
  - The only code he contributed was the Cold Start validation script
- It was helpful having at least one person who had actually read the task guidelines

Safety Net

- Started with a conservative baseline approach to which other things could be added
- Included a sanity checker
  - Could kill candidate values that don't meet the task specification (e.g., age: "old")
  - Could require values for some slots to be on a vetted lists
    - enumerations of religions, countries, etc...

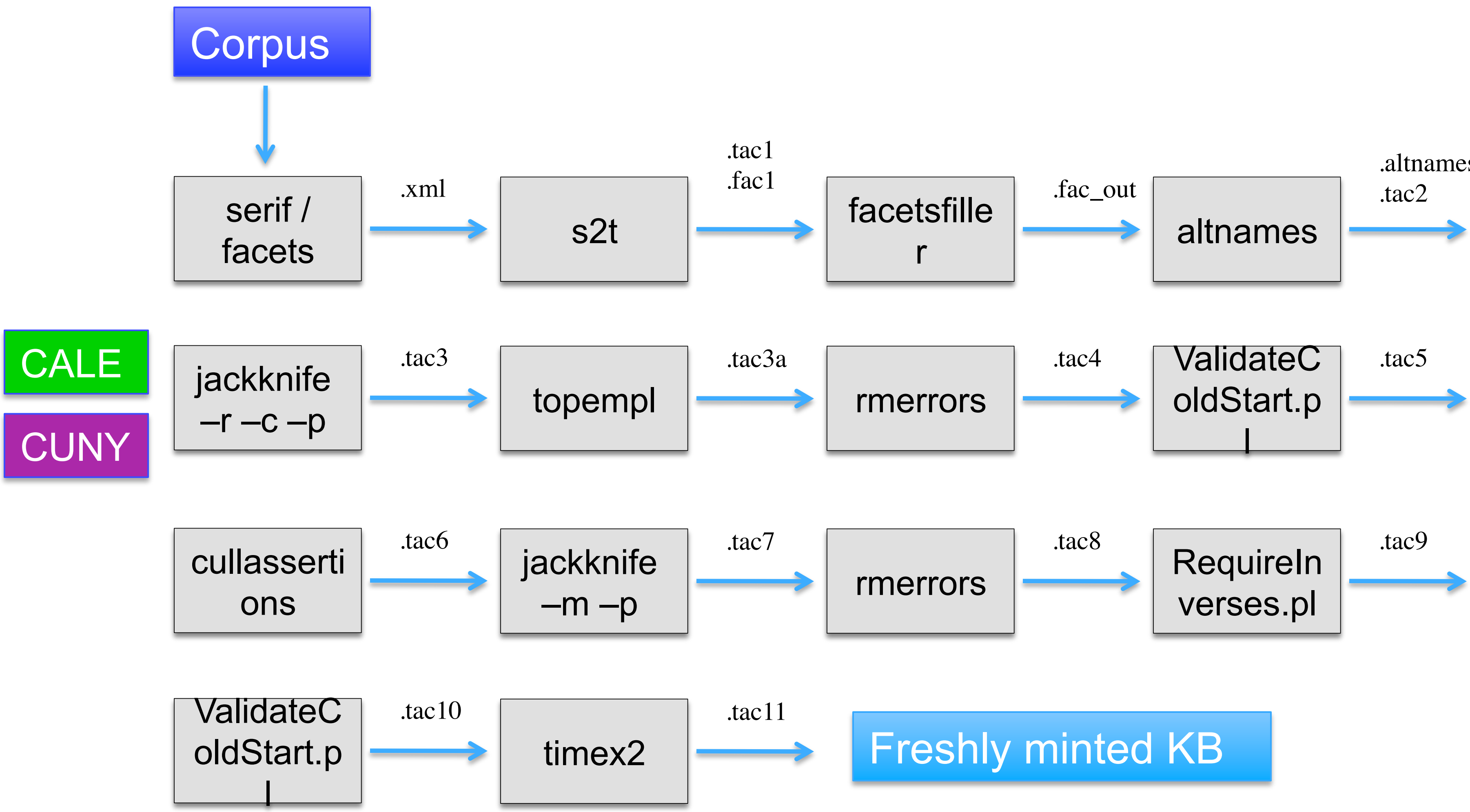


Minor Keys

- Glue
- Lack of Shame
  - My mantra was “lets just get model zero working first”
    - Coreference: string equality
    - Combination of multiple slot values: union, or most common for single value slots
    - Redundancy: arbitrary selection
- Tools
  - Some members of the team know how to write reusable code (KB format parsers, KB inspectors)
- Practice
  - We built ~10 test corpora to develop on.
    - Most were based on New York Times articles mentioning a string (e.g., “Reno” or “Baltimore Orioles”).
    - One is a subset of English Gigaword containing ~26k Washington Post articles from 2010

human language technology
center of excellence

Pipeline



human language technology
center of excellence

Component: SERIF

- BBN tool performs NIST/ACE task
  - Identify and classify named entities
  - perform intra-document coreference clustering
  - extract ACE relations between entities
  - detect certain event types
- ACE relations great micro-kernel for Cold Start
  - Relations are more generic than KBP
    - X PER-SOC Y (could be children, other, spouse)
    - X (PER) and Y (GPE.Nation) and relationship (GEN-AFF:Citizen-Resident-Religion-Ethnicity), then per:country_of_residence
  - Events like birth/death useful for some KBP slots
- Some issues
  - nested mentions, long named mentions

human language technology
center of excellence

Component: FACETS

- SERIF add-on tool which produces role/argument annotations for person noun phrases
- Maximum entropy classifier based on annotated noun phrases
- “52-year-old ambassador”
  - age, job title, diplomat role
- FACETS produces strings, not entities, so we tried to map string fills to document entities, when possible

Component: CUNY

- CUNY Slot Filling Toolkit was designed for slot filling, not Cold Start
- Created “queries” by working from SERIF NEs
  - Canonical mention from coref chain in each document
  - Tad computationally expensive, so we ran in parallel on COE cluster (1200 cores, Sun Grid Engine, ~4 TB RAM)
  - Only finds 9 KBP slot types
- CUNY slot fills are strings, not entities, so we tried to map string fills to document entities, when possible

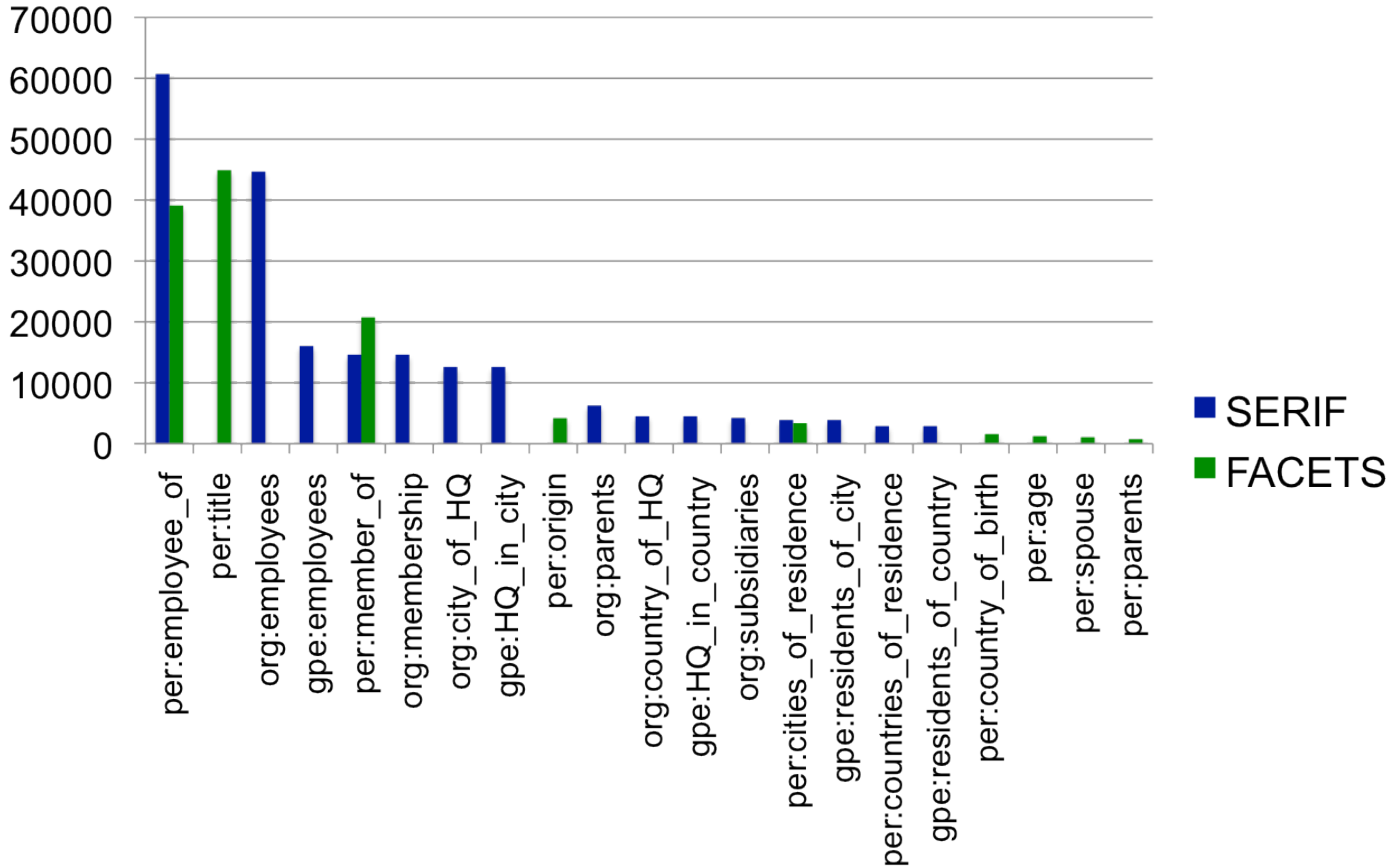
Component: CALE

- In addition to baseline (normalized string equality) for coreferene
- Used KBP KBIDs as entity ID when found in KBP reference KB
  - For NILs, used exact match NIL clustering
- Notably more effective for Cold Start
  - I did not foresee this, my presumption was that name polysemy wouldn't be a problem
  - Good coreference lets you correctly answer queries where slot fills come from documents other than where the focal entity occurs

Component: Light Inference

- Alternate names
  - If part of same SERIF coref cluster generate alternate_name slots
- Top Employees
  - If X employs Y in doc D, and Y has title (CEO|director|secretary|etc) in D
- Could probably do more in this regard, but at present the CS specification requires explicit evidence to assert slots

Practice: Washington Post articles



26k 2010 Washington Post articles (194k assertions)

Demo: Query Engine

- **Since no training data was available, we assessed our practice KBs qualitatively**

26k 2010 Washington Post articles (194k assertions)

human language technology
center of excellence

Good KELVIN

- Jared Fogle is an employee of Subway
- Freeman Hrabowski works for UMBC, founded the Meyerhoff Scholors Program, and attended both Hampton University and the University of Illinois
- Elena Kagan attended Oxford, Harvard, and Princeton
- The Applied Physics Laboratory is a subsidiary of Johns Hopkins University
- Southwest Airlines is headquartered in Texas

26k 2010 Washington Post articles (194k assertions)

human language technology
center of excellence

Bad KELVIN

- Harry Reid is an employee of the Republican Party. (KELVIN also learns he is employee of the Democratic Party)
  - Tend to see more trouble with popular entities
- Big Foot is an employee of Starbucks
- Steven Spielberg lives in Iran
- Jill Biden is married to Jill Biden

26k 2010 Washington Post articles (194k assertions)

1 year of WP = 1 computer scientist

- **KELVIN learns:**
  - ("ian soboroff”) per:employee_of ("national institute of standards and technology")
  - ("ian soboroff") per:title "computer scientist”

- paulmac@hpcc1: more WPB_ENG_20100506.0012.sgm
- One scholar Anderson has been talking with is Dan Cohen, the director of the Center for History and New Media at George Mason University; another is Ian Soboroff, a computer scientist with the National Institute of Standards and Technology, who studies how well technology works with human language. Soboroff hopes that in the future, Twitter can be used to help researchers understand how information was dispersed and what human networks in the early 21st century looked like.

human language technology
center of excellence

Experimental Results

	CUNY	CALE	0-hop slots P	0-hop slots R	0-hop slots F1	1-hop slots P	1-hop slots R	1-hop slots F1
hltcoe1			0.5256	0.2843	0.3690	0.1620	0.2182	0.1859
hltcoe2	YES		0.4929	0.3031	0.3753	0.1818	0.2406	0.2071
hltcoe3		YES	0.4865	0.5000	0.4932	0.1849	0.3510	0.2423
hltcoe4	YES	YES	0.4799	0.5155	0.4971	0.1842	0.4168	0.2555
hltcoe5	YES		0.4937	0.3053	0.3773	0.1453	0.2531	0.1846

CALE (better entity linking) materially improves recall

Precision (unsurprisingly) is lower for 0-hop vs. 1-hop

human language technology
center of excellence

Demo: KELVINPedia

26k 2010 Washington Post articles (194k assertions)

human language technology
center of excellence

Summary

- True "Cold Start" system: no direct use of external resources, live Internet, etc...
- Heavy leveraging of existing relation extraction, entity linking systems
- Over 50% relative gain in F-score when using CALE entity linker vs. exact-match baseline
- Best configuration:
  - Base + CUNY + CALE
- Now that we have a system, plan to exercise it on different data sets and see how it performs