

CSE 454  
Advanced Internet Systems

Project Logistics

Dan Weld

## Schedule

- Teams
  - Today
- Bidding
  - Starting tonight
- Meetings
  - Thurs / Fri

## Challenges

- Unfortunately for my slot, there is only one positive result in the test set, so I can't be very sure how well I am doing.

## Are Slot Values Really Rare?

- Gold-labeled data
- Test Data
- Actual NIST Challenge 2010
  - Training data
  - Queries “Joe Biden”
  - Union of all system's output (just on these queries)
  - Human checking
- 2011, 2012

## Challenges

- Raul Castro was born on June 3 , 1931 in Cuba 's Oriente Province and educated at Jesuit schools and the University of Havana .
  - My pattern catches \*Oriente\* but not \*Oriente Province\*.
- President Hugo Chavez was born *here* , his father has governed it for 10 years , and now his brother is seeking the governor's job .

## Challenges

- Hugo Chavez , was born on July 28 , 1954 , in Venezuela 's Sabaneta .
  - Because any amount of text can be in-between \*born\* and \*in Venezuela 's Sabaneta\* , it is hard to know if \*in\* is referring to \*born\*.
- Hugo Chavez was born before skinny dipping at a swimming pool in Venezuela 's Sabaneta.

## Org:Country

### 0. Find sentences with ORG

- Matching the string around an organization name to a list of pre-determined countries.
  - <country> company <org> ex. "U.S. company Disney"
  - <country>'s <org> ex. "Japan's Pentax Corp."
  - <org> of <country> ex. "Toyota of Japan"
  - <country>-based company <org> ex. "Japan-based company Honda"
- Matching the string around an organization name to a list of pre-determined U.S. states. U.S. organizations are probably very prevalent, and there are only 50 states, so we know that an organization from a particular state must also be from the U.S.
  - <org> of <city>,<state> ex. "Illinois Tool Works of Glenville, Illinois"
  - <org> in <city>,<state> ex. "North Phoenix Baptist Church in Arizona"
  - <state>-based company <org> ex. "N.J.-based Jackson Hewitt"
  - <state>'s <org> ex. "Washington's Microsoft"
- Matching within the organization for country names. Logically, an organization that has country name in its own name is probably from that country.
  - ex. "Myanmar Timber Enterprise"
- Matching within the organization for state names. This combines #2 and #3 from above.
  - ex. "Pennsylvania Supreme Court"

## Results

2010 scores:

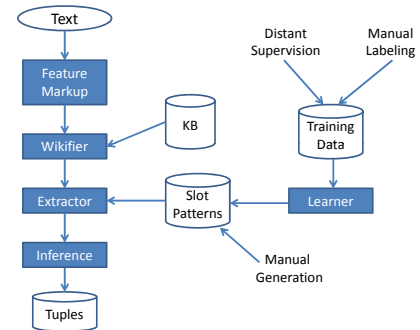
Recall: 5 / 23 = 0.17391305  
 Precision: 5 / 17 = 0.2  
 F1: 0.18604653 **0.29**

QueryID	Text that led to the assignment	Headquarters
SF223	National Union for the Total Independence of Angola	Angola
SF239	Myanmar Timber Enterprise	Myanmar
SF262	Taiwan's Noordhoff Craniofacial Foundation	Taiwan
SF233	Inter-American Press Association	America
SF224	Pennsylvania Supreme Court	U.S.
SF229	New Hampshire Institute of Politics	U.S.
SF230	Massachusetts House of Representatives	U.S.
SF236	Virginia-based National Military Family Association	U.S.
SF221	Paralyzed Veterans of America	America
SF232	Florida's Institute for Diversity and Ethics in Sport	U.S.
SF228	Professional Rodeo Cowboys Association in Colorado	U.S.
SF240	National Museum of Women in the Arts in Washington	U.S.
SF210	Britain's Crown Prosecution Service	Britain
SF263	Missouri-based National Christmas Tree Association	U.S.
SF256	Illinois Tool Works of Glenville, Illinois	U.S.
SF235	National Beef Packing Co. of Kansas	U.S.
SF265	Nitschmann Middle School in Bethlehem, Pennsylvania	U.S.
SF242	Japan's Pentax Corp.	Japan
SF234	N.J.-based Jackson Hewitt	U.S.
SF260	North Phoenix Baptist Church in Arizona	U.S.

## Org:Disolved

- In 1953, five years after the state was established the JNF was dissolved and re-organized as an Israeli company.
  - This should have been paired with the Jewish National Fund.
- My final solution, which is pretty terrible, doesn't use any regular expressions. It first eliminates totally incorrect answers by skipping ignored-slots and entities (non-"ORG" types). Then it checks if there's a date and an organization...Then it just checks if the word 'dissolved' appears anywhere in the sentence. If so, it links the org to the date.
  - Recall: 1/3 = 0.33333334
  - Precision: 1/50 = 0.02
  - F1: 0.03773585

## High-Level Architecture



## Slots

Person Slots		Organization Slots	
Slot	Type	Slot	Type
alternate names	Name List	alternate names	Name List
date of birth	Value Single	political/religious affiliation	Name List
age	Value Single	top members/employees	Name List
country of birth	Name Single	number of employees/members	Value Single
state/province of birth	Name Single	members	Name List
city of birth	Name Single	member of	Name List
origin	Name List	subsidiaries	Name List
date of death	Value Single	parents	Name List
country of death	Name Single	founded by	Name List
state/province of death	Name Single	founded	Value Single
city of death	Name Single	dissolved	Value Single
cause of death	String Single	country of headquarters	Name Single
countries of residence	Name List	state/province of headquarters	Name Single
state/provinces of residence	Name List	city of headquarters	Name Single
cities of residence	Name List	shareholders	Name List
schools attended	Name List	website	String Single
title	String List		
member of	Name List		
employee of	Name List		
religion	String Single		
spouse	Name List		
children	Name List		
parents	Name List		
siblings	Name List		
other family	Name List		
charges	Name List		

## Teams

- Named Entity Linking (1)
- Distant Supervision (1)
- InstaRead (1-2)
- Relation-Specific (3-5)
  - Time slots (bigger group?)
  - Location slots (city, state, country) – inference
  - Inverses: Parents/children/subsidiaries
  - List slots (normalization & de-dup)