human language technology
center of excellence

KELVIN 2.0: HLTCOE Progress in Cold Start Knowledge Base Population

19 November 2013

Paul McNamee, Tim Finin, Dawn Lawrie, & James Mayfield

human language technology
center of excellence

Talk Outline

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

- **Basic Pipeline**
- **2013 Undertakings**
  - **Cross Document Entity Coreference**
  - **Entity Consolidation**
  - **Inference Rules**
  - **Curation of Within-Document Coreference**
- **Experimental Results**
- **Demos**
  - **Browser**
  - **Query Engine**

human language technology
center of excellence

Basic Pipeline

1. Document Level Analysis
2. Cross-Document Coreference
3. Generate Inverses #1
4. Culling Assertions
5. Slot Value Consolidation
6. Inference
7. Inverses #2
8. Legal Cleanup

human language technology
center of excellence

BBN SERIF

- **NIST ACE tool**
  - Named entities, within-doc coref, relation extraction, some events
- **ACE relations**
  - 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
- **Issues**
  - Nested mentions, long named mentions, aggressive coreference

1. Document Analysis
2. Cross-Document Coref
3. Generate Inverses #1
4. Culling Assertions
5. Slot Value Consolidation
6. Inference
7. Inverses #2
8. Legal Cleanup

human language technology
center of excellence

BBN FACETS

- SERIF add-on tool
  - Produces role/argument annotations for person noun phrases
  - MaxEnt classifier based on annotated noun phrases
- "52-year-old ambassador"
  - age, job title, diplomat role
- FACETS yields strings, not entities
  - Must map string fills to document entities, when possible

1. Document Analysis
2. Cross-Document Coref
3. Generate Inverses #1
4. Culling Assertions
5. Slot Value Consolidation
6. Inference
7. Inverses #2
8. Legal Cleanup

human language technology
center of excellence

Cross Document Entity Coreference

- 2012: CALE
  - Entity Linker
  - Slow
- "Kripke"
  - Agglomerative Clusterer
  - No training
  - Precision-bias
  - "Faster"
- After this point, entities are "global" vs. "document"

1. Document Analysis
2. Cross-Document Coref
3. Generate Inverses #1
4. Culling Assertions
5. Slot Value Consolidation
6. Inference
7. Inverses #2
8. Legal Cleanup

human language technology
center of excellence

Inverses / sesrevnl

- **Relations are associated with specific KB entities**
  - **X per:spouse Y**
  - **X org:headquarters Y**

- **Producing inverses in our format simplifies our follow-on steps**

1. Document Analysis
2. Cross-Document Coref
3. Generate Inverses #1
4. Culling Assertions
5. Slot Value Consolidation
6. Inference
7. Inverses #2
8. Legal Cleanup

human language technology
center of excellence

Improving Precision and Recall

- **Culling Assertions**
  - Throw away bad looking slots
  - Candidate countries, states, religions not on gold-lists; non-numeric ages ("young")
- **Slot Consolidation**
  - Developing a unified view of an entity from single-document evidence
- **Inference**
  - More recall-enhancement vs. error detection
- **Legal Compliance**

1. Document Analysis
2. Cross-Document Coref
3. Generate Inverses #1
4. Culling Assertions
5. Slot Value Consolidation
6. Inference
7. Inverses #2
8. Legal Cleanup

human language technology
center of excellence

Kripke Cross-Document Coreference

- "Untrained" agglomerative clustering
  - All document entities start as singleton clusters
  - Combine if "good" name match and "good" context
  - Cascade of merging steps with relaxing constraints
  - No splits (currently)
  - Index of name variants per cluster kept at all times

- Series of name matching rules
  - Identical set of name mentions in C1 & C2 – score = 0.99
  - Two PER clusters with longest name having identical last word and all letters of one name appear in same order as name in other cluster = 0.74
    - Hilary Clinton and Hillary Diane Rodham Clinton
  - Highest scoring pattern wins

human language technology
center of excellence

Kripke Cross-Document Coreference

- **Context matching**
  - Uses co-occuring NEs as features
  - Take k=10 rarest NEs common to both clusters and sum normalized IDF to get a score between 0 and 10
  - Thus, “Mary” and “New York” are less discriminating than “Mike Huggins” and “Eau Claire”

- **Underconflation**
  - Can be hard to match prominent entities
  - Article about Pepsi using Facebook for marketing, but FB is mentioned tangentially
  - Well-known location (London, Tokyo) mentioned in non-focal article
    - For example, article about Michael Phelps may mention London briefly vs. say an article about the auto industry that mentions Tokyo, Toyota, Yokohama, Honda

human language technology
center of excellence

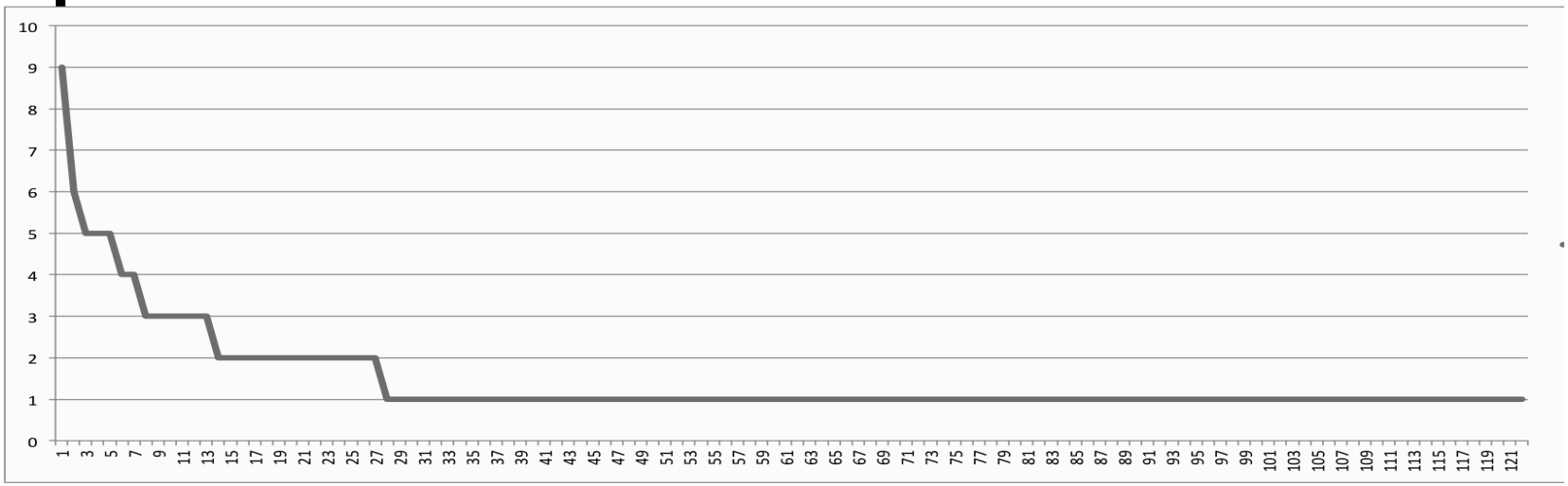
Slot Value Consolidation

- **Two tasks**
  - **For single-values slots (e.g., per:age) with more than one candidate value, select most likely one**
  - **For multi-valued slots (e.g., per:parents) with "too many" values, select best ones**
- **To paraphrase Tolstoy**
  - **All true facts are alike; each untrue fact is untrue in its own way.**
- **Basic strategy**
  - **Rank attested values based on their number of attesting documents**
  - **Rank inferred values lower than attested values**

Slot Value Consolidation

Examples from our 26K document Washington Post collection

- **We had 35 entities with two *per:age* values, 4 with three and 1 with four**
- **For *per:employee_of* we found entities with up to 201 values!**



Number of documents attesting the 122 values for per:employee_of for one entity

human language technology
center of excellence

Slot Value Consolidation

- **For single valued slots pick highest ranked candidate value**
- **Each multi-values slot has two thresholds**
  - **T1: accept nth ranked candidate if n ≤ T1**
  - **T2: accept nth ranked candidate if n ≤ T2 and has more than one attesting document**
- **Examples**

predicate	T1	T2
per:children	8	10
per:countries_of_residence	5	7
per:employee_of	8	10
per:religion	2	3
per:schools_attended	4	7
per:spouse	3	8

human language technology
center of excellence

Inference in Cold Start

- Kelvin 2.0 used procedures to implement forward chaining rules
- Benefits:
  - Greater efficiency
  - Use special algorithms – e.g., efficiently compute transitive closure of set of gpe:part_of relations
  - Easier to check provenance requirements
  - Easier to incorporate additional constraints via python code
- Drawbacks
  - Harder to write and read
  - Less portable
- Also, we extended our schema
  - per:sex, gpe:part-of

human language technology
center of excellence

Inference in Cold Start

- **Sound logical rules are derived from definitions**

  P1 per:parents P, P2 per:parents P

  => P1 per:siblings P2, P2 per:siblings P1

- **Plausible inference rules draw conclusions that are probably correct**

  P per:school_attended S, S city_of_headquarters C

  => P cities_of_residence C

- **Default rules draw conclusions that depend on the absence of contradictory evidence**

  P1 per:spouse P2, P1 per:sex "male", ~ P2 per:sex *

  => P1 per:sex "female"

human language technology
center of excellence

How many new facts?

- **Ran Kelvin on 26K Washington Post articles**
  - **Extracted 140,751 entities, ~2M facts**

- **The good news**
  - **Inference found 1,468,741 new relations**
  - **Note: GPE subsumption most productive**

- **The bad news**
  - **Most unusable because the entire relation not supported in a single document, e.g.:**
  - **P1 pre:parents P and P2 per:parents P read in different documents, so cannot assert P1 per:siblings P2**

human language technology
center of excellence

How many new facts?

NUMBER	%USABLE	PREDICATE
464472	5.1	org:stateorprovince_of_headquarters
358334	1.9	org:country_of_headquarters
244528	5.2	per:statesorprovinces_of_residence
188263	2.1	per:countries_of_residence
135991	0.0	gpe:part_of
16172	5.2	gpe:residents_of_stateorprovince
13926	100.0	per:top_member_employee_of
13926	100.0	org:top_members_employees
8794	7.6	per:stateorprovince_of_death
8038	5.2	per:stateorprovince_of_birth
6685	3.3	per:country_of_death
6107	2.1	per:country_of_birth
1561	100.0	per:employee_of
636	27.7	per:siblings
476	37.8	per:cities_of_residence
476	37.8	gpe:residents_of_city
356	58.4	per:other_family

26k 2010 WP articles

Inference Speed

- **Sibling**

X per:parents Z
Y per:parents Z
=>
X per:sibling Y
Y per:sibling X

Z per:children X
Z per:children Y
=>
X per:sibling Y
Y per:sibling X

- **Grandparents**

C per:parents B
B per:parents A
=>
C per:other_family A
A per:other_family C

B per:parents A
B per:children C
=>
C per:other_family A
A per:other_family C

human language technology
center of excellence

Good KELVIN

- Supreme Court Justice Elena Kagan attended Oxford, Harvard, and Princeton
- LeBron James is an NBA player for the NY Knicks, and he was born in Ohio
- The Applied Physics Laboratory is a subsidiary of Johns Hopkins University
- Southwest Airlines is headquartered in Texas
- The Smithsonian Institution has subsidiary organizations including: the Museum of Natural History, the American Art Museum, the National Museum of American History

26k 2010 Washington Post articles (194k assertions)

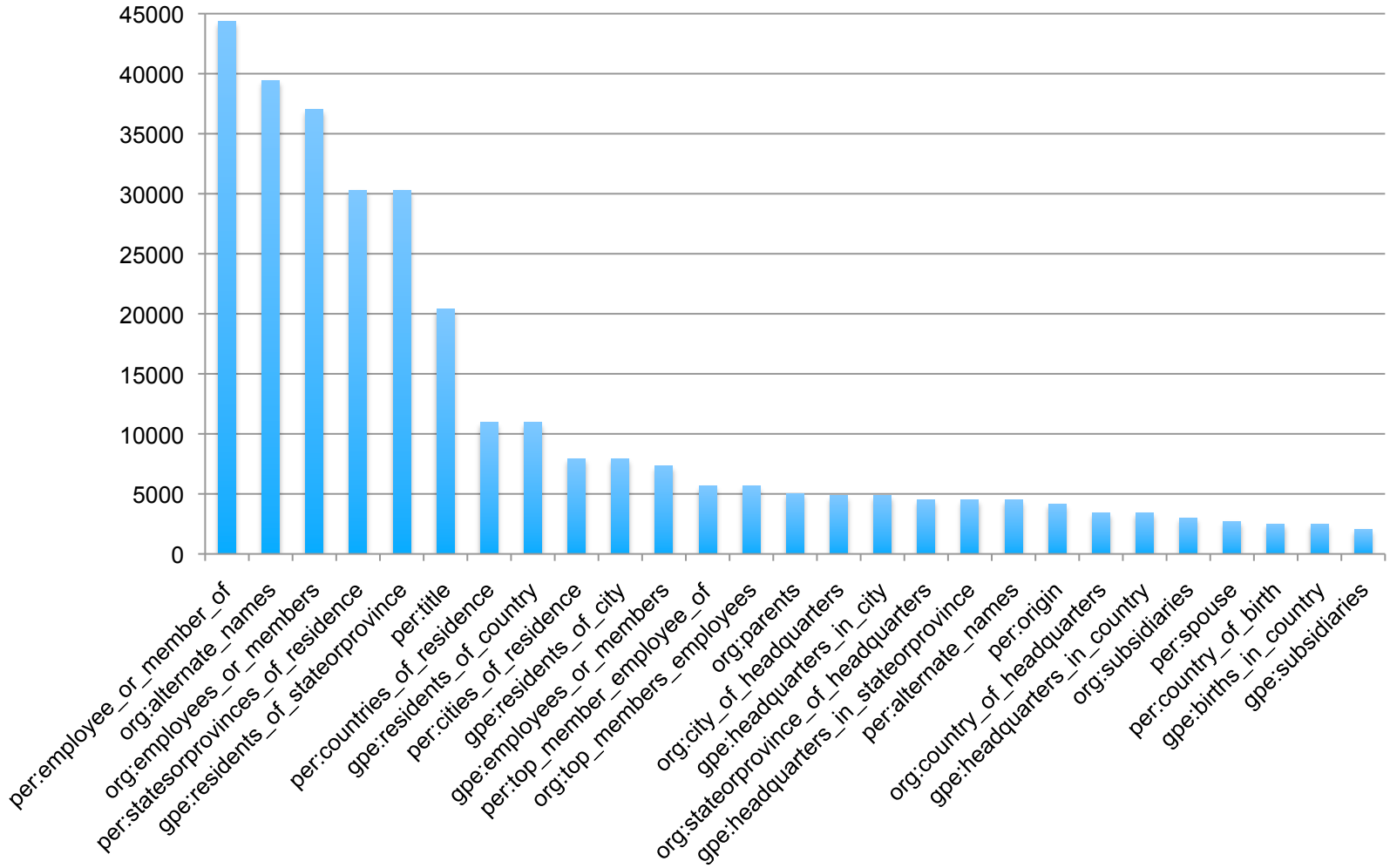
human language technology
center of excellence

Bad KELVIN

- Supreme Court Justice Sonia Sotomayor is a member of the US Senate
- Former US President Richard Nixon had a son, Charles
- Harry Reid is an member of the Republican Party. (KELVIN also learns he is member of the Democratic Party)
  - Tend to see more trouble with popular entities
- Steven Spielberg lives in Iran
- Jill Biden is married to Jill Biden

26k 2010 Washington Post articles (194k assertions)

Predicted Slot Prevalence



CS'13 Corpus – run with inference

Experiment 1: Better X-Doc Coref

		0-hop slots			1-hop slots		
	xdoc	P	R	F1	P	R	F1
hltcoe1	exact	0.429	0.267	0.329	0.072	0.109	0.087
hltcoe2	kripke	0.410	0.361	0.384	0.084	0.113	0.097

- **Compared "canonical mention exact match" vs. "Kripke"**
- **35% relative gain in 0-hop recall**
  - **But 2012 CALE system had ~50% gain in F1**

Experiment 2: Inference rules

			0-hop slots			1-hop slots		
	xdoc	infer	P	R	F1	P	R	F1
hltcoe2	kripke		0.410	0.361	0.384	0.084	0.113	0.097
hltcoe3	kripke	YES	0.350	0.278	0.310	0.082	0.124	0.098

- **Lower 0-hop precision and recall, but no difference with 1-hop queries**
  - **Problems due to provenance?**

human language technology
center of excellence

Billary, Brangelina, and Bennifer

KBID: 7654
Hillary Rodham Clinton
Sex: Female
Title: Senator
Title: Secretary of State
Born: Chicago, Ilinois

KBID: 4567
Bill Clinton
Sex: Male
Title: President
Born: Arkansas

"Bill and Hillary Clinton spent their vacation in the Hamptons... She was born in Chicago... Former President Clinton..."

KBID: 9999
Bill Clinton

AKA: Hillary Clinton
Sex: Female

Title: President

Born: Chicago, Ilinois

Experiment 3: Indoc Coref

				0-hop slots			1-hop slots		
	xdoc	infer	ndoc	P	R	F1	P	R	F1
hltcoe3	kripke	YES		0.350	0.278	0.310	0.082	0.124	0.098
hltcoe4	kripke	YES	YES	0.405	0.327	0.362	0.214	0.110	0.145

- **Detect and delete suspicious in-doc coref chains**
- **Improvements in precision and recall**
  - **Precision boost hoped-for**
  - **Recall boost unexpected, but could be due to aiding in xdoc coref decisions**
- **Gains transfer to 1-hop with higher precision**

Experiment 4: Corpus Augmentation

				0-hop slots			1-hop slots		
	xdoc	infer	xtra	P	R	F1	P	R	F1
hltcoe3	kripke	YES		0.350	0.278	0.310	0.082	0.124	0.098
hltcoe5	kripke	YES	YES	0.354	0.390	0.371	0.076	0.131	0.096

- **Additional documents used for learning outside the KBA collection**
  - **Random equal-sized sample of Gigaword articles**
- **Ultimately, all facts must be supported in CS corpus!**
- **Improvements possible due to:**
  - **Aiding in xdoc coref**
  - **Supporting inference**
  - **Different frequency of attestation**

human language technology
center of excellence

2013 Preliminary Results (Summary)

					0-hop slots			1-hop slots		
	xdoc	infer	ndoc	xtra	P	R	F1	P	R	F1
hltcoe1	exact				0.429	0.267	0.329	0.072	0.109	0.087
hltcoe2	kripke				**0.410**	0.361	**0.384**	0.084	0.113	0.097
hltcoe3	kripke	YES			0.350	0.278	0.310	0.082	0.124	0.098
hltcoe4	kripke	YES	YES		0.405	0.327	0.362	**0.214**	0.110	**0.145**
hltcoe5	kripke	YES		YES	0.354	**0.390**	0.371	0.076	**0.131**	0.096

- **Kripke coref helps with recall**
- **Inference *may* somewhat hurt**
  - **But with inference and other boosting techniques, performance is similar to our top hltcoe2 run**

human language technology
center of excellence

Miscellanea

- **Source code**
  - **Mainly Python & Java driven by csh scripts**
- **Cluster environment**
  - **Sun Grid Engine**
- **Run-times**
  - **24-48 hours**

human language technology
center of excellence

Demo: KELVINPedia

Anna Chapman (PER)

- Anna Chapman cities_of_residence New York City Unlike nine others arrested Monday, Chapman, 28, lived under her real name in New York City, the complaint said.
- Anna Chapman countries_of_residence Russian An anxious June 26 phone call from Russian spy Anna Chapman to her father, a KGB veteran working in Moscow's Ministry of Foreign Affairs, led the Obama administration to hasten the arrests the next day of Chapman and nine other "illegals" in the United States, according to U.S. law enforcement and intelligence sources.
- Anna Chapman employee_of Russian Anna Chapman, the Russian diplomat's daughter whose photos have become an Internet sensation, played with her red hair, attempting to tie it back.
- Anna Chapman employee_of Barclays Earlier, we learned that Anna Chapman, when she wasn't making sure there were plenty of pictures of herself in lingerie ready to fill the newspapers and blogs after her eventual capture, worked for Barclays in London.
- Anna Chapman origin "Russian" An anxious June 26 phone call from Russian spy Anna Chapman to her father, a KGB veteran working in Moscow's Ministry of Foreign Affairs, led the Obama administration to hasten the arrests the next day of Chapman and nine other "illegals" in the United States, according to U.S. law enforcement and intelligence sources.
- Anna Chapman parents Vasily Kushchenko Her father, Vasily Kushchenko, served in Kenya and has a senior position in the Ministry of Foreign Affairs, according the newspaper Komsomolskaya Pravda.
- Anna Chapman schools_attended Peoples' Friendship University of Russia International news sites have also gotten in on the action: A profile on the Russian site LifeNews.ru claims that Chapman is the daughter of the former Russian ambassador to Kenya, that she was raised by her grandmother, that she studied economics at the Peoples' Friendship University of Russia and once worked in banking.
- Anna Chapman statesorprovinces_of_residence Connecticut- In his initial phone call Saturday, June 26, he asked her to come to New York from Connecticut, where she was spending the weekend.

KB from 26k 2010 Washington Post articles (194k assertions)

human language technology
center of excellence

Demo: TripleStore Search

Endpoint : http://eb4.cs.umbc.edu:9090/ds/sparql

Output : Table

Configure request

```
PREFIX kbp: <http://tackbp.org/ontology/>
SELECT DISTINCT ?o ?name ?empl
{?o kbp:type kbp:ORG.
 $o kbp:country_of_headquarters ?x.
 ?o kbp:canonical_mention ?name.
 ?x kbp:canonical_mention $m.
 FILTER regex(?m, 'egypt', 'i').
 ?o kbp:employees ?y.
 ?y kbp:canonical_mention ?empl}
LIMIT 50
```

	o	name	empl
1	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood	Mohammed Akef
2	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood	Essam el-Erian
3	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood	Ayman Nour
4	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood	Mohammed Mursi
5	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood	Mohamed Mursi
6	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood party	Mohammed Akef
7	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood party	Essam el-Erian
8	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood party	Ayman Nour
9	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood party	Mohammed Mursi
10	http://tackbp.org/kb/e_WPB_ENG_20101205_0003_3	Muslim Brotherhood party	Mohamed Mursi
11	http://tackbp.org/kb/e_WPB_ENG_20100824_0060_30	EFG-Hermes Holding	Marise Ananian
12	http://tackbp.org/kb/e_WPB_ENG_20100824_0036_7	Supreme Council of Antiquities	Zahi Hawass
13	http://tackbp.org/kb/e_WPB_ENG_20100824_0036_7	Supreme Council for Antiquities	Zahi Hawass

KB from 26k 2010 Washington Post articles (194k assertions)

Property	Value
kbp:alternate_names	▪ Brotherhood ▪ Muslim Brotherhood
kbp:canonical_mention	▪ Muslim Brotherhood
kbp:canonical_name	▪ Muslim Brotherhood
kbp:city_of_headquarters	▪ kb:e_WPB_ENG_20100322_0001_2
kbp:co_mention	▪ kb:e_WPB_ENG_20100101_0022_11 ▪ kb:e_WPB_ENG_20100103_0010_11 ▪ kb:e_WPB_ENG_20100103_0023_37 ▪ kb:e_WPB_ENG_20100103_0024_11 ▪ kb:e_WPB_ENG_20100104_0008_4 ▪ kb:e_WPB_ENG_20100105_0074_23 ▪ kb:e_WPB_ENG_20100106_0023_13 ▪ kb:e_WPB_ENG_20100111_0028_10 ▪ kb:e_WPB_ENG_20100118_0030_4 ▪ kb:e_WPB_ENG_20100315_0007_14 ▪ kb:e_WPB_ENG_20100511_0110_14
kbp:communicated_with	▪ kb:e_WPB_ENG_20100118_0030_4
kbp:context	▪ A former member of the Muslim Brotherhood , Egypt 's Islamist opposition group , he decided to join the fight in Bosnia . ▪ Candidates affiliated with the Muslim Brotherhood became the largest opposition bloc in parliament in 2005 , winning 88 seats , roughly 20 percent . l ... »more» ▪ EGYPT -- CAIRO -- The new leader of the Muslim Brotherhood , Egypt 's biggest anti-government group , says his movement does n't oppose President Hosni Mubarak , a remark that may indicate the group 's retreat from an overt political role . ▪ Hamas also has close historical ties with the Muslim Brotherhood , Egypt 's biggest opposition group and is backed by Iran and Syria . ▪ He united most of the opposition behind him , including both the banned Muslim Brotherhood and liberal democrat Ayman Nour , who was imprisoned for challenging Mr. Mubarak in the 2005 presidential election . ▪ In Alexandria , Egypt 's second-largest city and a Brotherhood stronghold , opposition candidates described overt violations . l `` If they had simply ... »more»