Overview of Research at the IDIR Lab

Chengkai Li

Associate Professor, Department of Computer Science and Engineering

Director, Innovative Database and Information Systems Research (IDIR) Laboratory

University of Texas at Arlington

August 23, 2017

https://idir.uta.edu

Research at the Innovative Database and Information Systems Research Laboratory (IDIR)

Research areas

o Big Data and Data Science (Database, Data Mining, Natural Language Processing)

Theme of current research

o building large-scale human-assisting and human-assisted data and information systems with high usability, low cost and applications for social good

Research directions

o computational journalism, computational fact-checking
o crowdsourcing and human computation
o cybersecurity, social bots
o data exploration by ranking/skyline/preference queries
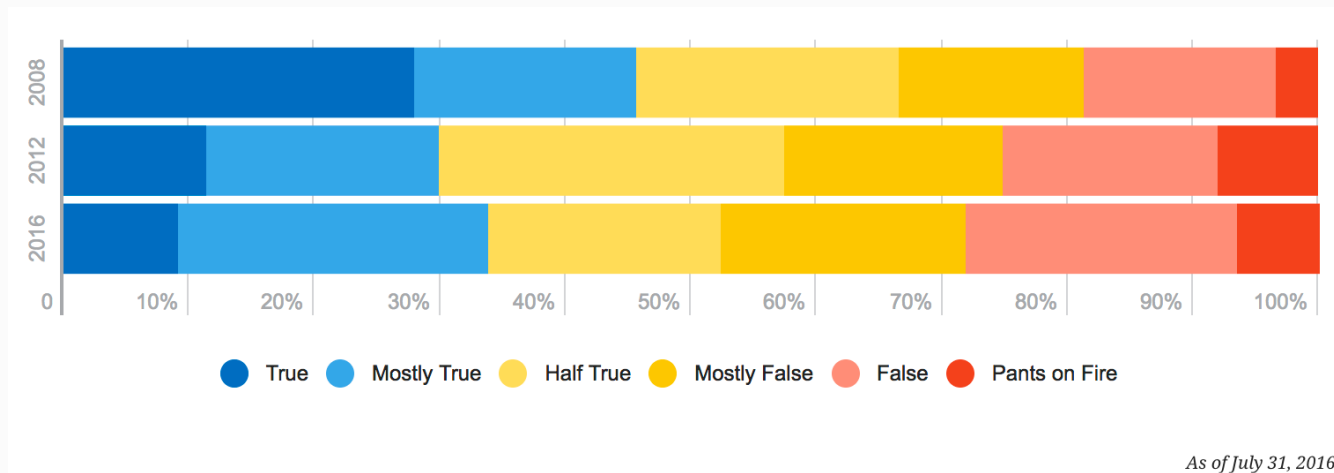o graph database usability, knowledge graphs

CLAIMBUSTER

Toward the Holy Grail of Automated Fact-checking

idir.uta.edu/claimbuster

Politicians Make False and Misleading Claims

As of July 31, 2016

Fact-checks on major party presidential nominees.
Source: PolitiFact.com

4

National Security Threats

o   Fake-news floods social media ("filter bubbles" and "echo chambers").

o Facebook trend topic algorithms promoted fake-news.

o A sample of 140,000 Twitter users in the battleground state of Michigan shared as many junk news items as professional news during the final ten days of the 2016 election. http://politicalbots.org/?p=1064

o Russian government interfered with the 2016 election. Fake-news websites and bots used.

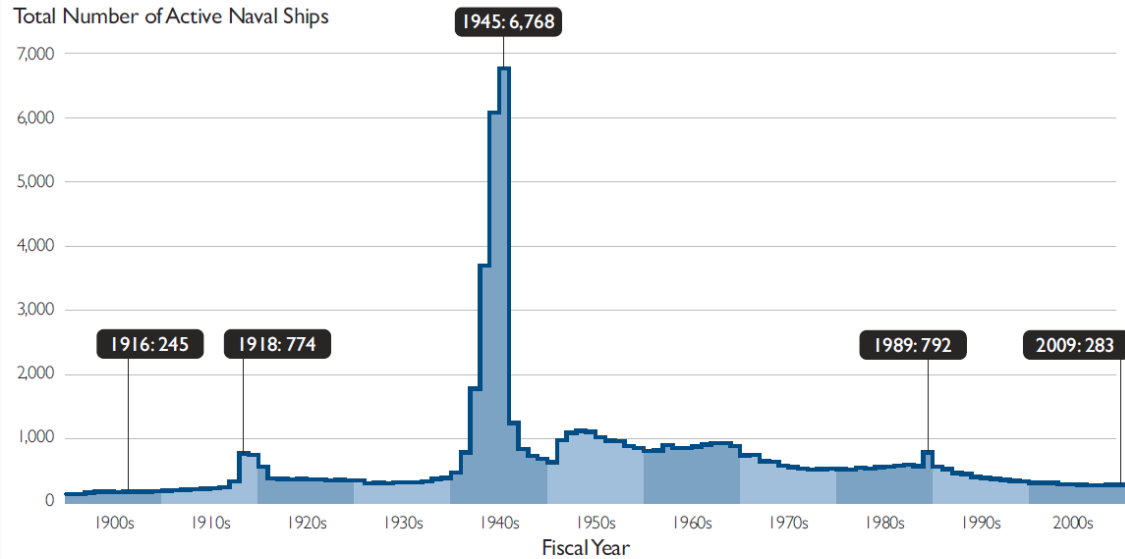o Pizzagate: conspiracy theory led to shooting

Combating Falsehoods

o 100+ active fact-checking sites in 2017 (PolitiFact.com, FullFact.org, CNN, Washington Post, New York Times, …)

o Google and Bing include fact-checks in search results.

o Facebook lets users report false items and flags items disputed by fact-checkers.

Fact-checking is Hard

"… our Navy is smaller than it's been since 1917", said Republican candidate Mitt Romney in third presidential debate in 2012.

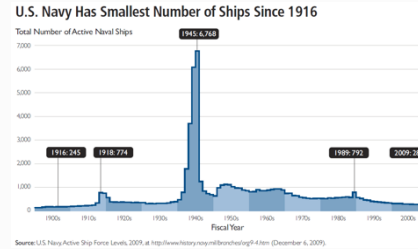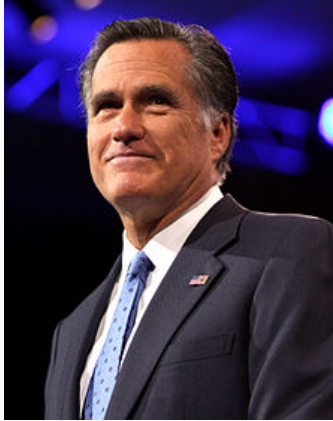## U.S. Navy Has Smallest Number of Ships Since 1916

Total Number of Active Naval Ships

1945: 6,768

1916: 245

1918: 774

1989: 792

2009: 283

Fiscal Year

Source: U.S. Navy, Active Ship Force Levels, 2009, at *http://www.history.navy.mil/branches/org9-4.htm* (December 6, 2009).

http://en.wikipedia.org/wiki/Mitt_Romney
http://s3.amazonaws.com/thf_media/2010/pdf/Military_chartbook.pdf

Fact-checking is Hard

"... our Navy is smaller than it's been since 1917", said Republican candidate Mitt Romney in third presidential debate in 2012.

U.S. Navy Has Smallest Number of Ships Since 1916

VS

Demanding Task vs. Lack of Automated Tools

Botminator: Automated Detection of Fake-news Bots

Objective
o discover social bots that publish and spread fake news

Recently funded by the Texas National Security Network
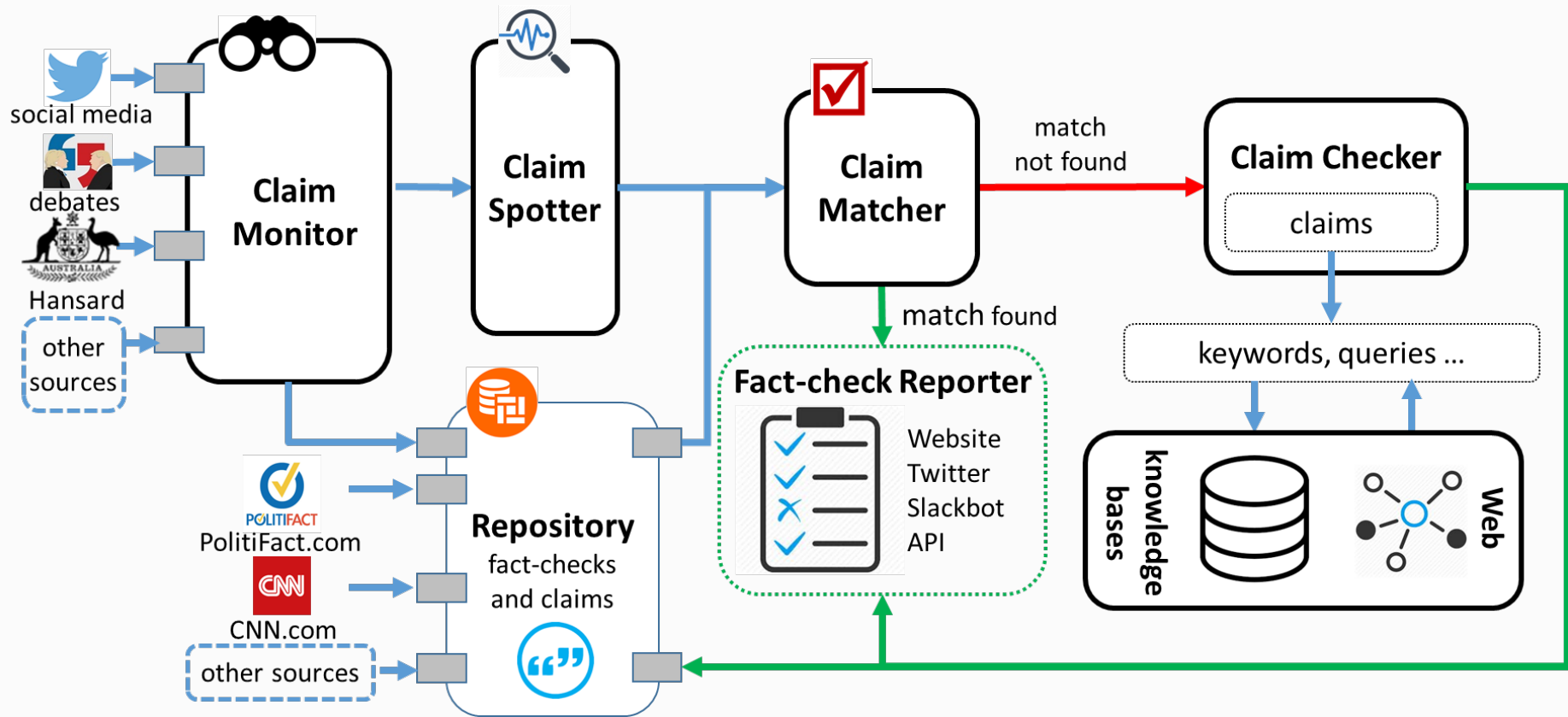
Interdisciplinary Approach
o Social network, graph, and text mining
o Code analysis
o News content and effect analysis

Collaborators on the Team
o Christoph Csallner (CSE, UTA)
o Mark Tremayne (Communication, UTA)
o Angela Lee (Emerging Media and Communication, UTD)
o Zhiqiang Lin (CS, UTD

ClaimBuster: The First-ever End-to-end System

Enter Your Own Text

2016 U.S. Presidential Debates

Hansard: Parliament of Australia

ClaimBuster API

ClaimBuster Slackbot

End-to-end Fact-checking (beta)

Tweets by @ClaimBusterTM

ClaimBuster Retweeted

Linda McMahon ✓
@SBALinda

The U.S. has 29 million small businesses & nearly 867,000 of them are in #Michigan. Almost ALL businesses in Michigan, 99.6% are #SmallBiz.

Aug 8, 2017

ClaimBuster Retweeted

Governor Walker ✓
@GovWalker

Since 2013, Kenosha County has seen nearly $1 billion in capital investment – creating over 6,000 jobs!

Embed   View on Twitter

ON AIR

2016 Third Presidential Debate Live. Oct. 19, 2016, 9 p.m. EST

Paticipants: Donald Trump, Hillary Clinton
Moderators: Chris Wallace

2016 Third Presidential Debate. Oct. 19, 2016, 9 p.m. EST

Paticipants: Hillary Clinton, Donald Trump
Moderators: Chris Wallace

2016 Second Presidential Debate. Oct. 9, 2016, 9 p.m. EST

Paticipants: Hillary Clinton, Donald Trump
Moderators: Martha Raddatz, Anderson Cooper, Chris Wallace

2016 Vice Presidential Debate. Oct. 4, 2016, 9 p.m. EST

Paticipants: Tim Kaine, Mike Pence, Hillary Clinton, Donald Trump
Moderators: Elaine Quijano, Martha Raddatz, Anderson Cooper

2016 First Presidential Debate. Sep. 26, 2016, 9 p.m. EST

Paticipants: Hillary Clinton, Donald Trump
Moderators: Lester Holt

2016 Democratic Party Presidential Debate. April 14, 2016, 9 p.m. EST

Paticipants: Bernard Sanders, Hillary Clinton, Donald Trump

13

2016 Third Presidential Debate. Oct. 19, 2016, 9 p.m. EST

Chronological Order | Order by Score

Least Check-worthy >=0.1 >=0.2 >=0.3 >=0.4 =0.5 >=0.6 >=0.7 >=0.8 >=0.9 >=1.0 Most Check-worthy

is tougher.   But they know what's going on.   They know it better than anybody.   They want strong borders.   They feel we have to have strong borders.   I was up in New Hampshire the other day.   The biggest complaint they have -- it's with all of the problems going on in the world, many of the problems caused by Hillary Clinton and by Barack Obama.   All of the problems -- the single biggest problem is heroin that pours across our southern border.   It's just pouring and destroying their youth.   It's poisoning the blood of their youth and plenty of other people.   We have to [Fact-check this] ders.   We have to keep the drugs out of our country.   We are -- right now, we're getting the drugs, they're getting the cash.   We need strong borders.   We need absolute -- we cannot give amnesty.   Now, I want to build the wall.   We need the wall.

Claim Checker - Knowledge Bases

Consulting the knowledge bases produced the following results:

Truth Rating Indeterminable

Question Asked
What is all of the problems-- the single biggest problem?

Response Recieved
The single biggest problem in communication is the illusion that it has ...

Claim Matcher

We found the following claims which have been professionally fact-checked. Check them out!

Truth Rating True

Claim
"Heroin .. pours across our southern borders."

Speaker Donald Trump

URL politifact

Truth Rating True

Claim Checker - Search Engine

We found the following information after processing some search engine results:

All of the problems -- the single biggest problem is heroin that pours across our southern border. It's just pouring and destroying their youth.

Similarity Rating 0.8320502943378437

URL source

"I was up in New Hampshire the other day," Trump said in the debate. "The biggest

GENIUS

FILTER BY   All Annotations

The biggest complaint they have -- it's with all of the problems going on in the world, many of the problems caused by Hillary Clinton and by Barack Obama
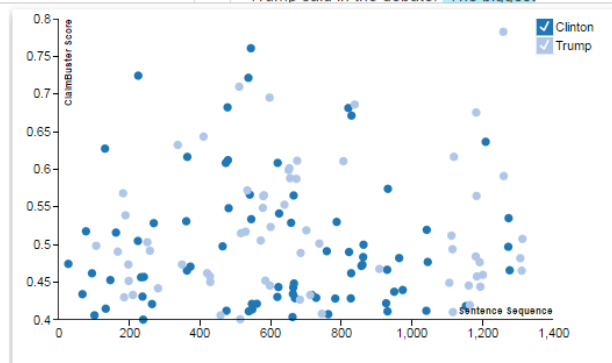
JosueCaraballo   17d

Is that true? What is the source?

report abuse

Upvote

Add a comment

ClaimBuster API Documentation

apidocs : ClaimBuster API

Show/Hide | List Operations | Expand Operations | Raw

GET	/all/{claim}	All
GET	/se/{claim}	Claim Checker - Search Engine
GET	/fm/{claim}	Claim Matcher
GET	/kb/{claim}	Claim Checker - Knowledge Bases
GET	/score_text/{text}	Claim Spotter - Get Score
GET	/score_url/{url}	Claim Spotter - Get Score URL

15

ClaimBuster APP 5:08 PM Only visible to you

Working on your request.

@gensheng

When my husband was president, we went from a $300 billion deficit to a $200 billion surplus and we were actually on the path to eliminating the national debt.

Score
0.657743014341

Working on your request.

@gensheng

We know how to get control of the debt.

Score
0.11061813488983

+ Message #slackbot_test_public ☺

ClaimBuster ➕ Slack

✳ ClaimBuster provides a platform for you to evaluate pieces of text and identify sentences that are check-worthy. This is our first step in our long-term goal of automating fact-checking. Our Slack app allows you to score individual sentences using the `/getscore claim` command, create a Dropbox folder on our servers using the `/createfolder e-mail` command, store text files from the web into your assigned Dropbox folder for future processing on our servers using the `/processfile e-mail file-url` command, and submit feedback to help us keep improving our services using the `/claimbusterfeedback message` command.

Add to **Slack**
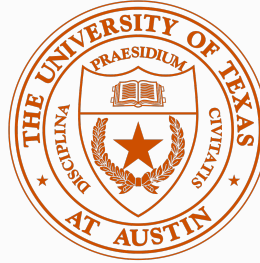
In the Press

Collaborators

Current/Former UTA Students

- Naeemul Hassan
- Gensheng Zhang
- Fatma Arslan
- Josue Caraballo

- Damian Jimenez
- Siddhantsanjeev Gawsane
- Shohedul Hasan
- Minumol Joseph

- Aaditya Kulkarni
- Anil Kumar Nayak
- Vikas Sable

Collaborators

- Bill Adair (Duke)
- Pankaj Agarwal (Duke)
- Christoph Csallner (UTA)
- Peter Frays (University of Technology Sydney)

- James Hamilton (Stanford)
- Angela Lee (UTD)
- Zhiqiang Lin (UTD)
- Mark Tremayne (UTA)

- You Wu (Google)
- Jun Yang (Duke)
- Cong Yu (Google)

Funding Sponsors

Finding Important Factual Claims:
A Supervised Learning Task

2016 Presidential Debates

Human Annotation

Feature Extraction

Learning Algorithm

Ground Truth

Feature Vectors

CLAIMBUSTER

Presidential Debate Transcripts (1960-2012)

Important factual claims
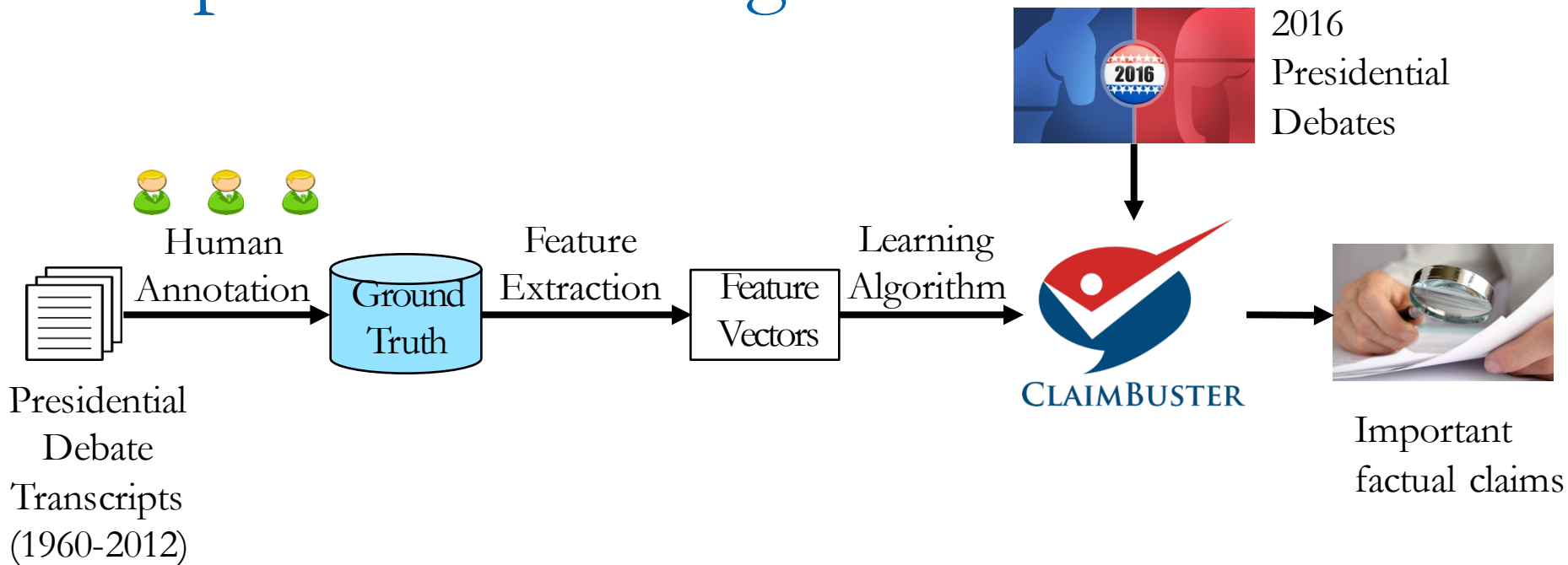
Classification and Ranking by Check-worthiness

CFS: Important factual claims

"We spend less on the military today than at any time in our history."   "The President's position on gay marriage has changed."   "More people are unemployed today than four years ago."

UFS: Unimportant factual claims

"I was in Iowa yesterday."   "My mother enjoys cooking."   "I ran for President once before."

NFS: No factual claims (opinions, questions & declarations)

"Iran must not get nuclear weapons."   "7% unemployment is too high."   "My opponent is wishy-washy."   "I will be tough on crime."   "Why should we do that?"   "Hello, New Hampshire!"   "Our plan is to reduce tax rate by 10%."

Feature Extraction

Word:

state, legislature, 87, percent, democrat

Sentiment: 0.032

I was in a <u>state</u> where my legislature was <u>87 percent</u> <u>Democrat.</u>

Part-of-Speech:

Noun

Entity Type:

Quantity

Concept:

United States
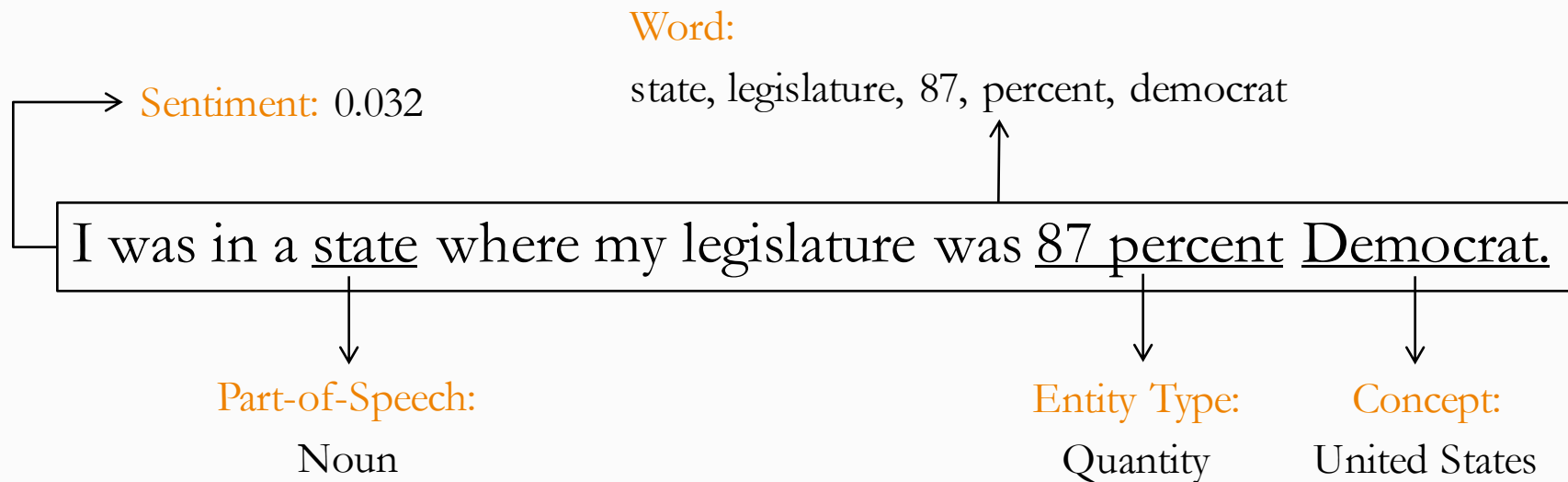
Sentiment: [-1.0 to 1.0]

Word: tf-idf scores of 6130 words (excluding rare words)

POS Tag: 43 tags

Entity Type: 26 types

Feature Importance

o 6615 features in total

o Used a Random Forest Classifier to calculate the importance
  of each feature.
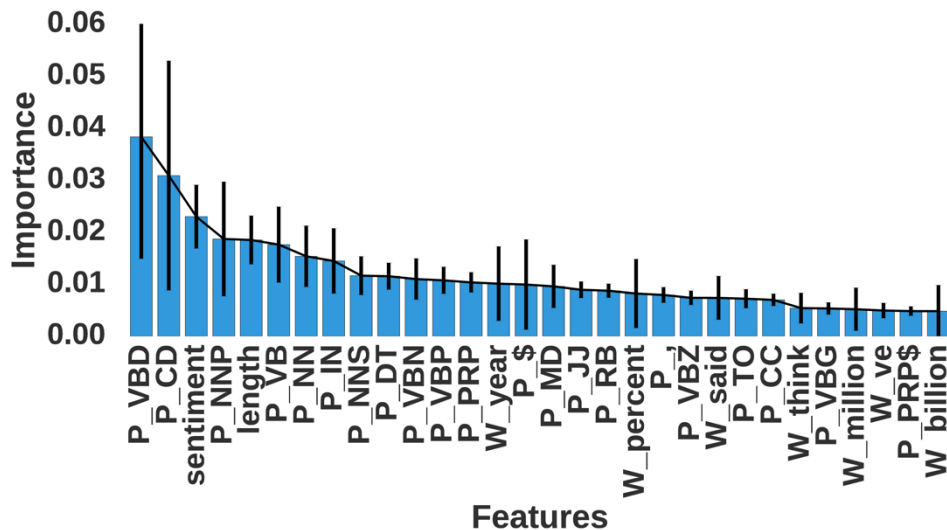
Word	percent; people; jobs
POS Tag	noun; cardinal number; past tense; preposition
Entity Type	Quantity; Country; FieldTerminology; Person
Concept	United States Senate; Barack Obama
Sentiment	

Our Own CrowdFlower

Coder Quality

$$R_p = \frac{L_p}{L}^{1.5} * \left(3 - \frac{7 * LQ_p}{0.2}\right) * 0.6^{\frac{|SKIP_p|}{|ANS_p|}}$$

Ground Truth Collection
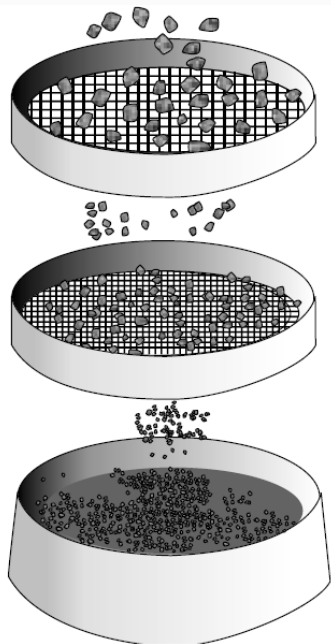
20788 sentences

374 coders

76552 labels

86 top-quality coders

52333 labels

Majority voting

20617 admitted sentences

- 20788 sentences
- 20 months, 374 coders, ~$4,000 paid
- 30 training sentences
- 1032 screening sentences (731 NFS, 63 UFS, 238 CFS) to detect spammers & low-quality coders

Class	Count	Percentage
CFS	4849	23.52%
UFS	2097	10.17%
NFS	13671	66.31%

http://www.engineersdaily.com/2014/03/basics-of-soil-compaction.html

Case Study: 2016 U.S. Presidential Election Debates

Table 6: Score differences between sentences fact-checked and those not chosen for checking

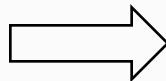| Platforms	avg(YES)	avg(NO)	t-value	p-value
CNN	0.433	0.258	21.137	1.815E-098
PolitiFact	0.438	0.258	16.362	6.303E-060
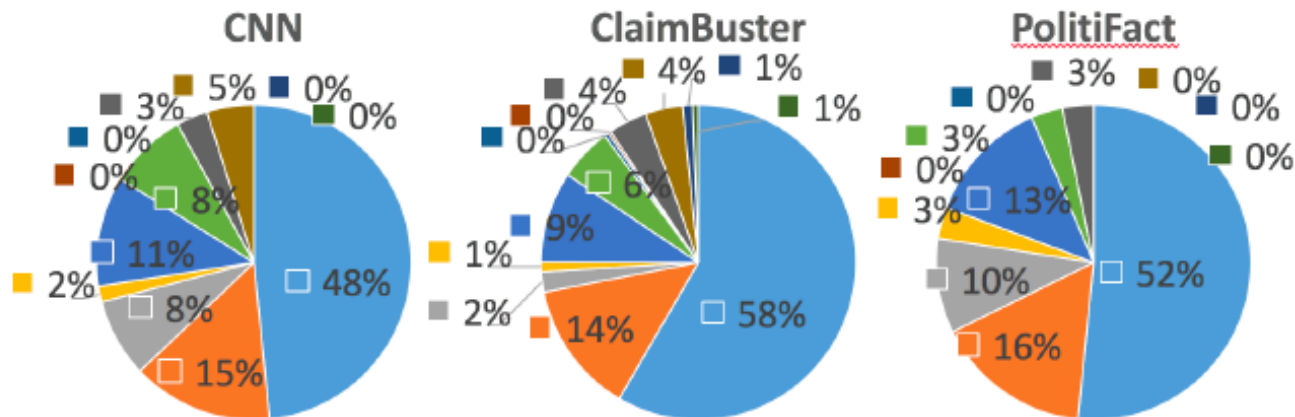
Figure 14: Comparison of topic distributions of CNN, PolitiFact fact-checked sentences and sentences scored high (≥ 0.5) by ClaimBuster

Other Projects and Demos

- Demonstration Videos https://vimeo.com/channels/1024406

- ClaimBuster (idir.uta.edu/claimbuster) Automated, Live Fact-Checking
- CrewScout (idir.uta.edu/crewscout) Expert Team Finding by Skyline Groups
- ERQ: Entity-Relationship Query (idir.uta.edu/erq) Structured Entity-Centric Query on Wikipedia
- Facetedpedia (idir.uta.edu/facetedpedia) Faceted search interface for Wikipedia
- FactWatcher (idir.uta.edu/factwatcher) Automated Monitoring of Facts from Real-World Events
- GQBE (idir.uta.edu/gqbe) Graph Query by Example
- Orion (idir.uta.edu/orion) Auto-suggestion for Interactive Graph Query Formulation
- TableView Generating Preview Tables for Knowledge Graphs

FactWatcher: Story-finding

Finding and monitoring number-based facts pertinent to real-world events. The facts are leads to news stories.
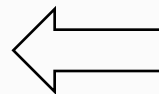
id	player	day	month	season	team	opp_team	pts	ast	reb
t_1	Bogues	11	Feb.	1991-92	Hornets	Hawks	4	12	5
t_2	Seikaly	13	Feb.	1991-92	Heat	Hawks	24	5	15
t_3	Sherman	7	Dec.	1993-94	Celtics	Nets	13	13	5
t_4	Wesley	4	Feb.	1994-95	Celtics	Nets	2	5	2
t_5	Wesley	5	Feb.	1994-95	Celtics	Timberwolves	3	5	3
t_6	Strictland	3	Jan.	1995-96	Blazers	Celtics	27	18	8
t_7	Wesley	25	Feb.	1995-96	Celtics	Nets	12	13	5

Real-world event (sports, transportation, crime, weather, finance, social media)
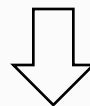
Find facts

Wesley had 12 points, 13 assists and 5 rebounds on February 25, 1996 to become the first player with a 12/13/5 (points/assists/rebounds) in February.

Generate factual claim

http://en.wikipedia.org/wiki/Basketball

FactWatcher
idir.uta.edu/factwatcher

Excellent Demo Award

Orion

Prototype

http://idir.uta.edu/orion

Introduction Video

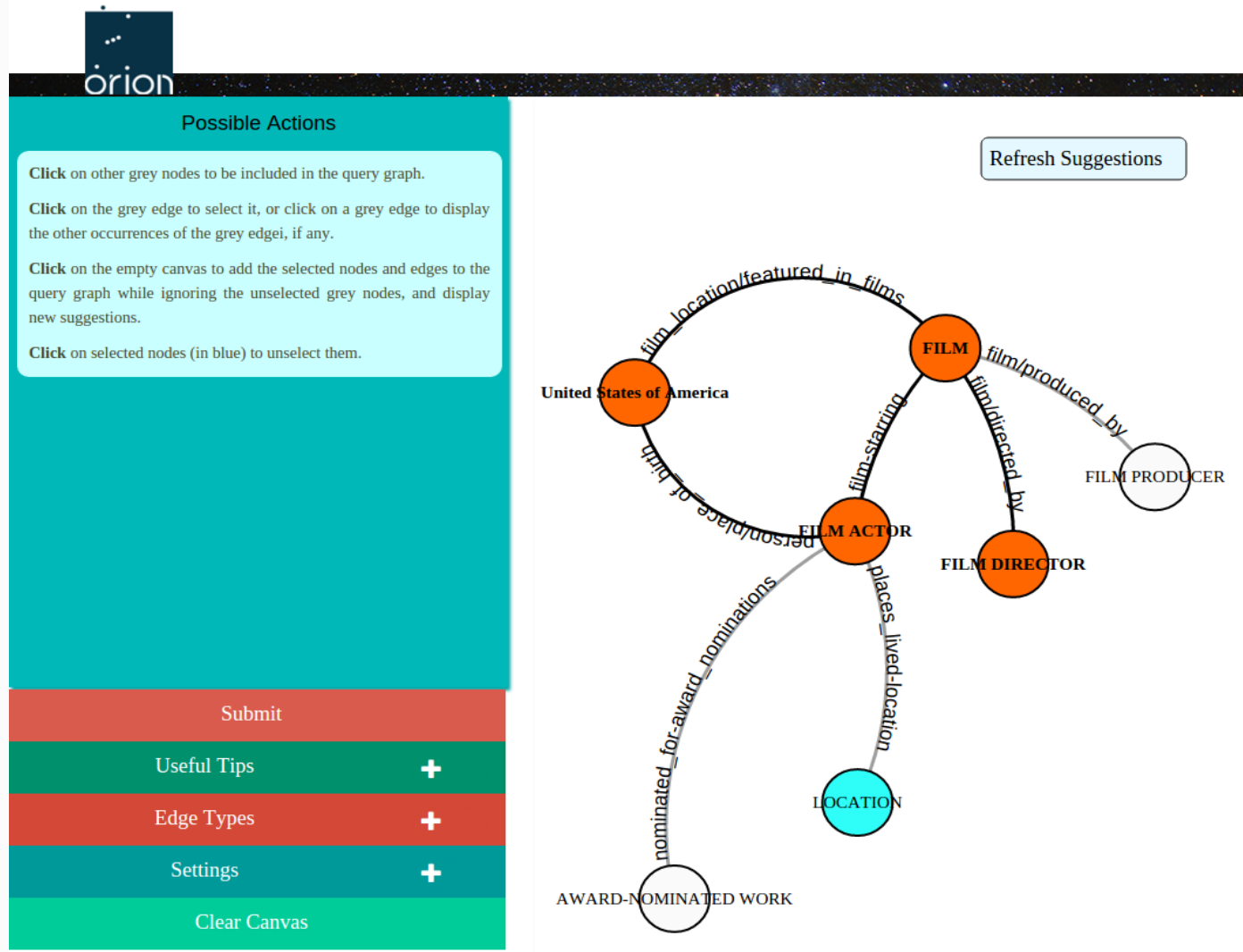http://bit.ly/1O0GnNo

Orion GUI

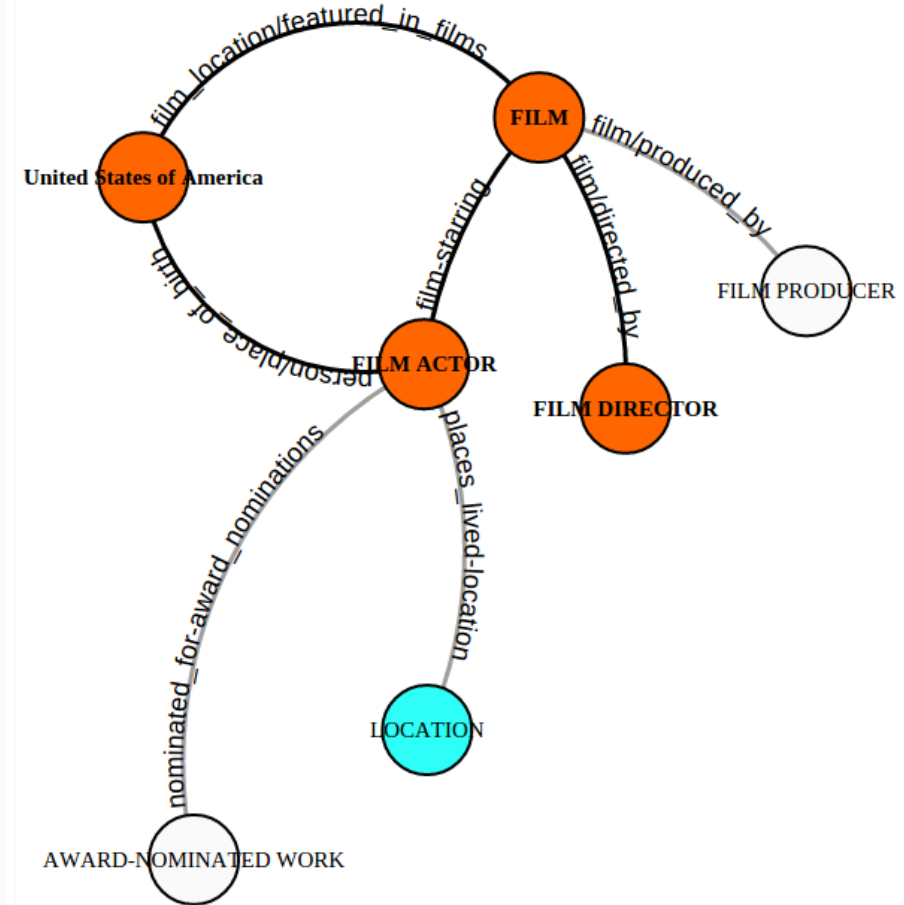Dynamic list of all possible user actions at any given moment

Control panel for various settings and tips

Active Mode

Grey edges and nodes automatically suggested in **active mode:**

o   Accepted by user (blue): positive edges
o   Ignored by user: negative edges

Passive Mode

A new node added in **passive mode**

A new edge added in **passive mode**

Publications

- Cross-lingual Entity Alignment via Joint Attribute-Preserving Embedding. Zequn Sun, Wei Hu, and Chengkai Li. ISWC 2017.

- ClaimBuster: The First-ever End-to-end Fact-checking System. Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. VLDB 2017, 1945-1948. Demonstration description.

- Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. KDD 2017, pages 1803-1812.

- Computational Fact Checking through Query Perturbations. You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. TODS, 42(1): 4:1-4:41, 2017.

- Graph Querying Meets HCI: State of the Art and Future Directions. Sourav S. Bhowmick, Byron Choi, Chengkai Li. SIGMOD 2017, 1731-1736. Tutorial.

- Querying knowledge Graphs By Example entity tuples. Nandish Jayaram, Arijit Khan, Chengkai Li, Xifeng Yan, Ramez Elmasri. ICDE 2016: 1494-1495.

Publications (cont'd)

o Comparing Automated Factual Claim Detection Against Judgments of Journalism Organizations. Naeemul Hassan, Mark Tremayne, Fatma Arslan, and Chengkai Li. 2016 Computation+Journalism Symposium.

o Generating Preview Tables for Entity Graphs. Ning Yan, Sona Hasani, Abolfazl Asudeh, Chengkai Li. SIGMOD 2016: 1797-1811. (**Most Reproducible Paper Award**)

o Detecting Check-worthy Factual Claims in Presidential Debates. Naeemul Hassan, Chengkai Li, Mark Tremayne. CIKM 2015, pages 1835-1838.

o Crowdsourcing Pareto-Optimal Object Finding by Pairwise Comparisons. Abolfazl Asudeh, Gensheng Zhang, Naeemul Hassan, Chengkai Li, and Gergely Zaruba. CIKM 2015, pages 753-762.

o The Quest to Automate Fact-Checking. Naeemul Hassan, Bill Adair, James Hamilton, Chengkai Li, Mark Tremayne, Jun Yang and Cong Yu. 2015 Computation+Journalism Symposium.

o Online Frequent Episode Mining. Xiang Ao, Ping Luo, Chengkai Li, Fuzhen Zhuang, and Qing He. ICDE 2015, pages 891-902.

Publications (cont'd)

- VIIQ: Auto-Suggestion Enabled Visual Interface for Interactive Graph Query Formulation. Nandish Jayaram, Sidharth Goyal, Chengkai Li. PVLDB, 8(12): 1940-1943, August 2015. Demonstration description.

- Querying knowledge graphs by example entity tuples. Nandish Jayaram, Arijit Khan, Chengkai Li, Xifeng Yan, Ramez Elmasri. TKDE, 27(10): 2797-2811, October 2015.

- GQBE: Querying knowledge graphs by example entity tuples. Nandish Jayaram, Mahesh Gupta, Arijit Khan, Chengkai Li, Xifeng Yan, Ramez Elmasri. ICDE, pages 1250-1253, 2014. Demonstration description.

- Data In, Fact Out: Automated Monitoring of Facts by FactWatcher. Naeemul Hassan, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. VLDB 2014, pages 1557-1560. Demonstration description. (**excellent demonstration award**)

- Finding, Monitoring, and Checking Claims Computationally Based on Structured Data. Brett Walenz, You (Will) Wu, Seokhyun (Alex) Song, Emre Sonmez, Eric Wu, Kevin Wu, Pankaj K. Agarwal, Jun Yang, Naeemul Hassan, Afroza Sultana, Gensheng Zhang, Chengkai Li, Cong Yu. 2014 Computation+Journalism Symposium.

Publications (cont'd)

- Toward Computational Fact-Checking. You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. VLDB 2014, pages 589-600.

- iCheck: computationally combating "lies, d-ned lies, and statistics". You Wu, Brett Walenz, Peggy Li, Andrew Shim, Emre Sonmez, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. SIGMOD 2014, pages 1063-1066.

- Incremental Discovery of Prominent Situational Facts. Afroza Sultana, Naeemul Hassan, Chengkai Li, Jun Yang, Cong Yu. ICDE 2014, pages 112-123.

- Discovering General Prominent Streaks in Sequence Data. Gensheng Zhang, Xiao Jiang, Ping Luo, Min Wang, Chengkai Li. ACM TKDD, 8(2):article 9, June 2014.

- Discovering and Learning Sensational Episodes of News Events. Xiang Ao, Ping Luo, Chengkai Li, Fuzhen Zhuang, Qing He, and Zhongzhi Shi. WWW 2014, pages 217-218.

- On "One of the Few" Objects. You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. KDD 2012, pages 1487-1495.

Publications (cont'd)

- Prominent Streak Discovery in Sequence Data. Xiao Jiang, Chengkai Li, Ping Luo, Min Wang, Yong Yu. KDD 2011, pages 1280-1288.

- Computational Journalism: A Call to Arms to Database Researchers. Sarah Cohen, Chengkai Li, Jun Yang, Cong Yu. CIDR 2011, pages 148-151. (**3rd place in best Outrageous Ideas and Vision (OIV) Track paper competition**)