Evaluating AIDA using hand annotated data

Aman Madaan

February 5, 2014

Abstract

Aim of this excercise was to find out how AIDA performs with hand annotated data. Label for the data were obtained with AIDA and ratio of number of annotations made by AIDA to the number of annotations in the labeled data for a particular entity is used as the metric to evalutate the success. The mean value of 1.12 indicates that AIDA did pretty well for entities that were recogized by both AIDA and CSAW. However, the recall is quite low, 0.2606, $\frac{989}{3795}$. Low recall is partially explained by a large number of extraneous tags in CSAW data.

1 Setup

1.1 Labelled data

Labeled data was taken from `http://www.cse.iitb.ac.in/soumen/doc/CSAW/Annot/`

- 104 annotated files
- 3795 entities
- 19000 spots

The annotations are available in XML format. A sample annotation is as follows :

```
 <annotation>
<docName>ganeshTestDoc.txt</docName>
<userId>amitsingh</userId>
<wikiName>Sachin Tendulkar</wikiName>
<offset>420</offset>
<length>9</length>
</annotation>
```

1.2 Benchmarking AIDA with CSAW data

To compare how AIDA performs, we obtained annotations from AIDA in the same XML schema as provided by CSAW team. The userId given was AIDA.

```
 <annotation>
<docName>ganeshTestDoc.txt</docName>
<userId>AIDA</userId>
<wikiName>Mike Denness</wikiName>
<offset>17</offset>
<length>12</length>
</annotation>
```

1.3 Annotated files

XML file containing annotations for 104 files as above and for other 560 files is available at `www.cse.iitb.ac.in/~amanmadaan/store`

1.4 Code

The source code (without jar files) is hosted at `http://github.com/madaan/aida_benchmark`

2 Annotation Statistics

Total entities annotated by AIDA	989
Total entities annotated by CSAW Team	3795
Total entities common to both	483
CSAW Entities missed by AIDA : (full list follows)	3312
AIDA Entities missed by CSAW : (full list follows)	506

Table 1: Statistics on Entities

n	483
min	0.03333333333333333
max	9.0
$mean$	1.121289147465606
std dev	0.9020144035396026
$median$	1.0
$skewness$	4.657584815674553
$kurtosis$	28.2250405390826

Table 2: Statistics for $score = \frac{\text{Annotations by AIDA}}{\text{Annotations by CSAW}}$

3 List of entities Annotated by both CSAW and AIDA

The score is defined as $score = \frac{\text{Annotations by AIDA}}{\text{Annotations by CSAW}}$

Entity Name	CSAW Count	AIDA Count	Ratio
Habib Beye	1	1	1.0
Dursley	6	4	0.6666666666666666
New Orleans Hornets	1	2	2.0
Joe Biden	4	2	0.5
Luc Montagnier	1	6	6.0
World Wide Web	1	5	5.0
Duke University	1	1	1.0
California Air Resources Board	1	1	1.0
Lehman Brothers	2	1	0.5
Albert Einstein College of Medicine	1	1	1.0
Republic of China	2	1	0.5
Evander Holyfield	3	6	2.0
Sachin Tendulkar	9	10	1.1111111111111112
Barry Hearn	1	1	1.0
Richard Garriott	3	17	5.666666666666667
Minsk	2	2	1.0
Ernesto Bertarelli	4	2	0.5
Thaksin Shinawatra	1	1	1.0
Norway	4	6	1.5
Washington Mutual	1	1	1.0
Bangalore	2	1	0.5
Nelson Piquet	2	2	1.0
University of Pittsburgh	1	1	1.0
Florida International University	1	1	1.0
Dagbladet	1	1	1.0
Mexico	6	5	0.8333333333333334
Kabul	10	8	0.8
Honda	2	2	1.0
Rajasthan	1	1	1.0
Joey Barton	1	1	1.0
Pakistan	15	14	0.9333333333333333
Simone Bolelli	1	3	3.0
Human papillomavirus	26	2	0.07692307692307693
General Growth Properties	2	1	0.5
David Beckham	2	2	1.0
National Basketball Association	5	6	1.2
Cervarix	1	1	1.0
Abdullah of Saudi Arabia	1	1	1.0
United States Environmental Protection Agency	1	1	1.0
McGill University Health Centre	4	2	0.5
Cornell University	4	3	0.75
Netherlands national football team	1	4	4.0
National Indoor Arena	1	1	1.0

Sarah Palin	8	6	0.75
Roseland Ballroom	1	1	1.0
Georgia Institute of Technology	5	4	0.8
Obafemi Martins	1	1	1.0
Nobel Prize	9	7	0.7777777777777778
Benazir Bhutto	4	4	1.0
Jonathan Zittrain	4	4	1.0
Afghanistan	53	74	1.3962264150943395
Mississippi	1	2	2.0
Purdue University	1	1	1.0
FC Barcelona	1	1	1.0
Mazda	9	9	1.0
Argonne National Laboratory	1	1	1.0
Robinho	1	1	1.0
Taiwan	1	1	1.0
Sitamarhi	1	1	1.0
Midfielder	5	1	0.2
Vienna	1	1	1.0
Croatia	5	5	1.0
Aaron Diamond AIDS Research Center	1	1	1.0
Muzaffarpur	1	1	1.0
Ronaldinho	1	1	1.0
Patna	1	1	1.0
Thomas Huckle Weller	1	1	1.0
Bert van Marwijk	1	1	1.0
The New York Times	2	2	1.0
Charlie Patton	1	1	1.0
Larry Ellison	5	4	0.8
Bernard Kouchner	3	3	1.0
Rio Ferdinand	2	2	1.0
Regions Financial Corporation	4	2	0.5
Jan Andersson	2	1	0.5
Joe Kinnear	1	3	3.0
Eritrea	1	1	1.0
Steven Gerrard	1	7	7.0
Spain	1	4	4.0
Nikolai Valuev	3	3	1.0
Russell Martin	2	2	1.0
Matt Kemp	1	1	1.0
Hillary Rodham Clinton	2	2	1.0
Haiti	2	3	1.5
France	7	5	0.7142857142857143
Dana White	1	1	1.0
Dow Jones Industrial Average	1	2	2.0
Bear Stearns	1	2	2.0
Facebook	2	2	1.0
Mike Denness	1	1	1.0
Chase Utley	1	1	1.0
Nomar Garciaparra	1	2	2.0
Thailand	4	3	0.75

International Security Assistance Force	1	1	1.0
Frederick Chapman Robbins	1	1	1.0
Ralph Stanley	2	2	1.0
Topper Headon	1	1	1.0
Virender Sehwag	3	3	1.0
Joel Lamangan	2	2	1.0
Cristiano Ronaldo	3	8	2.6666666666666665
Bluetooth	1	1	1.0
Kevin Keegan	1	1	1.0
Nad Ali District	2	1	0.5
Credit Suisse	2	1	0.5
Davis Cup	1	1	1.0
Bank of America	51	26	0.5098039215686274
Michael Bisping	4	14	3.5
George Harrison	1	1	1.0
GE Global Research	2	1	0.5
UEFA Champions League	2	5	2.5
Nicolas Sarkozy	3	3	1.0
Stuttgart	1	1	1.0
Uzbekistan	1	1	1.0
Zabul Province	7	6	0.8571428571428571
Antonio Villaraigosa	1	1	1.0
Leslie Gonda	2	1	0.5
Methylnaltrexone	4	3	0.75
Jamie Dimon	2	1	0.5
Nathaniel Parker	1	1	1.0
Walter Dix	2	1	0.5
Tom Moody	1	1	1.0
Sweden	2	1	0.5
Owen Hargreaves	1	1	1.0
Myriad Genetics	2	1	0.5
Olivier Rochus	1	1	1.0
San Francisco Chronicle	1	1	1.0
Boston Celtics	1	2	2.0
Barclays Capital	2	1	0.5
Michael Fincke	1	1	1.0
Bournville	1	1	1.0
Anthony Fauci	1	2	2.0
The Seattle Times	2	1	0.5
Carnegie Mellon University	2	2	1.0
Roger Federer	9	9	1.0
Columbia University	4	2	0.5
Ontario	3	3	1.0
Greenberg	1	3	3.0
Wayne Rooney	6	6	1.0
Brown University	2	3	1.5
IBM	6	2	0.3333333333333333
United States Department of the Treasury	1	3	3.0
Jamaica	2	2	1.0
Islamabad	2	2	1.0

Berlin	1	1	1.0
Sandia National Laboratories	3	2	0.6666666666666666
Natural Resources Defense Council	1	2	2.0
David Haye	1	1	1.0
Blake DeWitt	2	3	1.5
Academia Sinica	1	1	1.0
Rahul Dravid	3	3	1.0
Emile Heskey	3	6	2.0
Islamism	1	1	1.0
Manchester	1	1	1.0
Christopher Denise	1	1	1.0
Belgium	2	1	0.5
Pasteur Institute	1	1	1.0
Reuters	15	7	0.4666666666666667
Chile	9	3	0.3333333333333333
Ranibizumab	1	1	1.0
Latin America	7	4	0.5714285714285714
Hydrocodone	2	1	0.5
Chrissie Hynde	5	5	1.0
Matthew Upson	1	1	1.0
Islamic Republic News Agency	1	1	1.0
John Ruiz	1	1	1.0
England national football team	8	2	0.25
Barack Obama	6	4	0.6666666666666666
David Nalbandian	2	2	1.0
Gilles Simon	1	1	1.0
Vietnam	2	1	0.5
Urdu	1	1	1.0
Cory Wade	1	1	1.0
Baldomero Olivera	1	1	1.0
Thomson Financial	2	1	0.5
Shea Stadium	4	1	0.25
Louisiana	1	2	2.0
Andy Roddick	2	2	1.0
Eleuthera	2	2	1.0
Dick Cheney	1	1	1.0
Steve Waugh	1	1	1.0
Luis von Ahn	1	1	1.0
The Bahamas	24	26	1.0833333333333333
United States Department of Defense	2	2	1.0
Joe Strummer	1	1	1.0
Chennai	1	1	1.0
Nobel Prize in Physiology or Medicine	1	1	1.0
Tommy Robredo	2	2	1.0
Stony Brook University	1	4	4.0
South Coast Air Quality Management District	4	4	1.0
Sudan	2	2	1.0
Darren Lehmann	1	1	1.0
Green fluorescent protein	32	11	0.34375
Chesapeake Energy	4	1	0.25

Ernests Gulbis	1	1	1.0
California	6	2	0.3333333333333333
Mardy Fish	1	1	1.0
Central Bank of The Bahamas	1	1	1.0
Saudi Arabia	14	6	0.42857142857142855
Never Ending Tour	1	1	1.0
Arizona	2	2	1.0
Alaska	3	1	0.3333333333333333
India	15	19	1.2666666666666666
Qari Mohammad Yousuf	1	1	1.0
Wembley Stadium	2	1	0.5
Fabio Capello	13	12	0.9230769230769231
Epigallocatechin gallate	6	5	0.8333333333333334
Kazakhstan	5	5	1.0
Philippines	1	1	1.0
Bernd Schuster	5	5	1.0
Iceland	4	3	0.75
Felix Wankel	1	1	1.0
University of Michigan	3	3	1.0
Russia	5	4	0.8
Moksha	1	1	1.0
Iran	2	2	1.0
Rafael Furcal	2	2	1.0
The Arctic Challenge	3	1	0.3333333333333333
Iraq	3	5	1.6666666666666667
Texas	1	2	2.0
Nancy Pelosi	6	6	1.0
Nicky Cook	1	1	1.0
Polyphenol	2	1	0.5
Wells Fargo	10	6	0.6
Donald Bradman	1	1	1.0
Stockholm Open	4	4	1.0
New York University	1	1	1.0
Argentina	1	1	1.0
Dirk Kuyt	5	2	0.4
Los Angeles Dodgers	5	5	1.0
Broad Institute	1	1	1.0
Ethiopia	1	1	1.0
United States Secretary of Defense	1	1	1.0
Turing test	1	1	1.0
Jun Lana	1	2	2.0
Forbes	2	2	1.0
Arctic	3	3	1.0
Wilfried Sauerland	3	3	1.0
Carl Lewis	1	1	1.0
AIDS	13	12	0.9230769230769231
Nick Van Exel	1	1	1.0
Steve Hodge	1	9	9.0
University of Kansas	1	1	1.0
International Space Station	7	6	0.8571428571428571

Europe	3	4	1.3333333333333333
Wasim Akram	1	1	1.0
Abu Rawash	1	1	1.0
JPMorgan Chase	10	5	0.5
Frank Lampard	1	6	6.0
University of Utah	1	1	1.0
Taliban	35	25	0.7142857142857143
Peru	4	1	0.25
Simon Gray	1	1	1.0
GlaxoSmithKline	1	1	1.0
Seychelles	1	1	1.0
Geomatics	1	1	1.0
Portugal	1	1	1.0
Jeev Milkha Singh	3	1	0.3333333333333333
Bihar	4	5	1.25
Andre Ethier	1	2	2.0
La Liga	1	2	2.0
Flavio Briatore	2	2	1.0
DNA sequencing	3	13	4.333333333333333
German language	1	1	1.0
Steve Darcis	1	1	1.0
Shawn Crawford	2	1	0.5
Alexandra Hospital	1	1	1.0
Bill Ayers	1	1	1.0
Fernando Alonso	4	7	1.75
John Arne Riise	1	1	1.0
Christopher Timothy	1	1	1.0
Don Quarrie	1	1	1.0
Belarus	5	5	1.0
Andhra Pradesh	1	2	2.0
Stanford University	3	3	1.0
BBC	4	2	0.5
New York City	15	3	0.2
Avery Johnson	1	1	1.0
University of Akron	1	1	1.0
Elvis Presley	1	1	1.0
Roche Applied Science	2	1	0.5
Rafael Nadal	2	2	1.0
Christian Malcolm	2	2	1.0
University of Texas at Austin	1	1	1.0
Harvard University	2	1	0.5
Lahore	2	2	1.0
Massachusetts Institute of Technology	3	3	1.0
Kazakhs	1	1	1.0
Jyoti Randhawa	1	1	1.0
Sunshine Dizon	15	5	0.3333333333333333
Pat Burrell	1	1	1.0
Hajipur	2	2	1.0
University of Connecticut	2	1	0.5
Ricky Hatton	2	2	1.0

Zinedine Zidane	1	1	1.0
Stefan Schumacher	13	4	0.3076923076923077
Associated Press	17	8	0.47058823529411764
Randy Lerner	1	1	1.0
The Huffington Post	1	1	1.0
Metin Sitti	1	1	1.0
World Boxing Organization	2	1	0.5
Jermain Defoe	1	1	1.0
Osama bin Laden	2	1	0.5
Visceral leishmaniasis	17	1	0.058823529411764705
Elias Zerhouni	2	1	0.5
AOL	1	1	1.0
Jim Keltner	1	1	1.0
Usain Bolt	1	4	4.0
Chris Leben	1	1	1.0
Citigroup	11	8	0.7272727272727273
Deutsche Bank	3	2	0.6666666666666666
John Terry	5	5	1.0
Barnett Shale	1	1	1.0
Italy	5	2	0.4
North America	1	1	1.0
Trieste	3	2	0.6666666666666666
Scotland	3	3	1.0
Celsius	3	3	1.0
Small Press Expo	1	1	1.0
Terai	1	1	1.0
Charles Keating	1	8	8.0
Igor Andreev	1	1	1.0
Helmand Province	2	1	0.5
Steve Jobs	2	1	0.5
DAX	1	1	1.0
CAPTCHA	22	1	0.045454545454545456
Diabetes mellitus	30	1	0.03333333333333333
Carl Froch	1	1	1.0
Romford	1	1	1.0
Germany	10	9	0.9
George Foreman	1	1	1.0
Tajikistan	1	1	1.0
United States dollar	1	1	1.0
Candlewick Press	5	5	1.0
Oslo	3	3	1.0
National Science Foundation	1	1	1.0
Mark Viduka	1	1	1.0
Hampden Park	1	1	1.0
Wakil Ahmed Muttawakil	2	3	1.5
Kyrgyzstan	1	1	1.0
Djibouti	1	1	1.0
Amphotericin B	2	1	0.5
Greg Chappell	1	1	1.0
Windows Live	1	1	1.0

Internet Explorer	1	1	1.0
Federal Reserve System	2	3	1.5
Egypt	1	1	1.0
New Zealand	1	1	1.0
Shane Victorino	2	3	1.5
Gardasil	2	2	1.0
University of Dayton	1	1	1.0
Chris Waddle	1	1	1.0
Beijing	7	6	0.8571428571428571
Slimbridge	2	2	1.0
National Assembly	2	2	1.0
Shahid Afridi	1	1	1.0
Mohali	1	1	1.0
American Cancer Society	1	1	1.0
Sultan Ibragimov	1	1	1.0
Karolinska Institutet	1	1	1.0
International Lease Finance Corporation	2	1	0.5
Michael Bevan	1	1	1.0
Moscow	1	1	1.0
University of Washington	2	2	1.0
Switzerland	7	3	0.42857142857142855
London	2	10	5.0
Utah	2	1	0.5
Hamid Karzai	3	2	0.6666666666666666
Skype	1	1	1.0
Bob Dylan	23	23	1.0
Africa	6	1	0.16666666666666666
Julia Ward Howe	3	3	1.0
National Cancer Institute	2	2	1.0
Brian Lara	2	2	1.0
Gareth Barry	2	2	1.0
Budapest	1	1	1.0
Clitheroe	1	1	1.0
Gerry Ritz	1	1	1.0
Jean Pascal	1	1	1.0
Michael Moorer	1	1	1.0
Singapore	3	3	1.0
Bruce Springsteen	1	1	1.0
United Nations	4	4	1.0
Maternal bond	1	2	2.0
Intel Corporation	4	4	1.0
Hubert Ingraham	5	5	1.0
Hiroki Kuroda	9	9	1.0
Ashley Cole	1	1	1.0
Cameron White	1	1	1.0
Christopher Kimball	9	5	0.5555555555555556
Microsoft	5	5	1.0
UEFA	1	1	1.0
Nobel Prize in Chemistry	2	1	0.5
Carnival Cruise Lines	2	1	0.5

Damien Martyn	1	1	1.0
Howard Hughes Medical Institute	1	1	1.0
Pedro Feliz	1	1	1.0
Hawaii	1	1	1.0
Digital Equipment Corporation	1	1	1.0
Karachi	2	1	0.5
James Posey	6	6	1.0
BBC Sport	2	2	1.0
England	15	23	1.5333333333333334
Cuba	1	1	1.0
Ben Bernanke	1	1	1.0
New York	6	15	2.5
Shoe Carnival	2	1	0.5
Nepal	2	2	1.0
Nevada	1	1	1.0
Wallace Spearmon	1	1	1.0
Advanced Cell Technology	1	1	1.0
Katrina Halili	2	2	1.0
Central Intelligence Agency	1	1	1.0
Mohammed Omar	2	1	0.5
Theo Walcott	4	4	1.0
John McCain	15	13	0.8666666666666667
North Eleuthera	1	1	1.0
Fuji Speedway	1	1	1.0
Wachovia	9	7	0.7777777777777778
Ultimate Fighting Championship	22	18	0.8181818181818182
United States	95	42	0.4421052631578947
Unix	1	1	1.0
Andreas Seppi	2	2	1.0
Felipe Massa	4	5	1.25
Robert Gates	8	8	1.0
MSN	1	1	1.0
Iker Casillas	1	1	1.0
Lewis Hamilton	2	4	2.0
Brazil	9	8	0.8888888888888888
Formula One	1	2	2.0
Australia	9	6	0.6666666666666666
Brooklyn	1	1	1.0
Churandy Martina	1	1	1.0
Bangladesh	1	1	1.0
Shoaib Malik	2	5	2.5
ING Group	2	1	0.5
Steven Rose	1	1	1.0
The Clash	1	3	3.0
Paris	4	4	1.0
Saskatchewan	1	2	2.0
Martin Chambers	1	1	1.0
Forrester Research	1	1	1.0
Chris Paul	1	1	1.0
Pakistan Peoples Party	1	1	1.0

Philadelphia Phillies	7	7	1.0
South Africa	1	1	1.0
Afghan National Army	2	1	0.5
Xiamen	1	1	1.0
Microsoft Windows	9	1	0.1111111111111111
Bischoff	1	1	1.0
Rockefeller University	1	1	1.0
Joe Calzaghe	2	4	2.0
Thomas Reh	2	3	1.5
Jamie Moyer	2	5	2.5
Marat Safin	1	1	1.0
Quezon City	2	2	1.0
Stockholm	1	1	1.0
Tokyo	3	3	1.0
Bell Labs	1	1	1.0
Canada	2	6	3.0
Welsh people	1	1	1.0
Daniel Lanois	2	4	2.0
Cheltenham	1	1	1.0
Robert Kubica	3	3	1.0
Public Broadcasting Service	4	1	0.25
Alan Turing	1	1	1.0
Jonas Salk	4	3	0.75
New York Army National Guard	1	1	1.0
Samyama	3	2	0.6666666666666666
Kenya	1	1	1.0
NATO	10	8	0.8
Mayo Clinic	1	1	1.0
NASA	8	8	1.0
Google	12	21	1.75
Czech Republic	1	1	1.0
Mozilla Firefox	1	1	1.0
Sialkot	11	11	1.0
Yemen	1	1	1.0
American International Group	9	8	0.8888888888888888
Victoria Hale	1	1	1.0
Merrill Lynch	5	5	1.0
Keith Harris	1	1	1.0
The Pentagon	2	2	1.0
Sourav Ganguly	1	2	2.0