Evaluating AIDA using hand annotated data

Aman Madaan

February 6, 2014

Abstract

Aim of this exercise was to benchmark AIDA by obtaining entity annotations on an already (hand) annotated dataset. A low recall is observed, caused by failure to resolve ambiguitites correctly, and partially because of a large number of extraneous hand annotated tags. For entities on which both the datasets agree, it is observed that for about 62% entities, the number of annotations exactly match. For about 87% of the entities, the number of annotations is within ± 5% of the real value.

1 Setup

1.1 Labeled data

The ground truth 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 Getting annotations

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>
```

Both the annotation files were then read into a hashmap of the form, $h(Entity) - > Count$. Maps for the two files were then scanned to generate the statistics presented in the next section.

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 Results

Various statistics on the ratio of number of annotations made by AIDA to the number of annotations in the labeled data for a particular entity are used as a metric of evaluation. 2 different setups were tried but results obtained were almost the same.

2.1 Setup 1

AIDA was run in *FastLocalDisambiguation* mode. The Corpus used was yago trained from 2010 version of Wikipedia. A total of 104 documents were annotated.

Total entities annotated by AIDA	989
Total entities annotated by CSAW Team	3795
Total entities common to both	548
CSAW Entities missed by AIDA : (full list follows)	3247
AIDA Entities missed by CSAW : (full list follows)	441

Table 1: Statistics on Entities

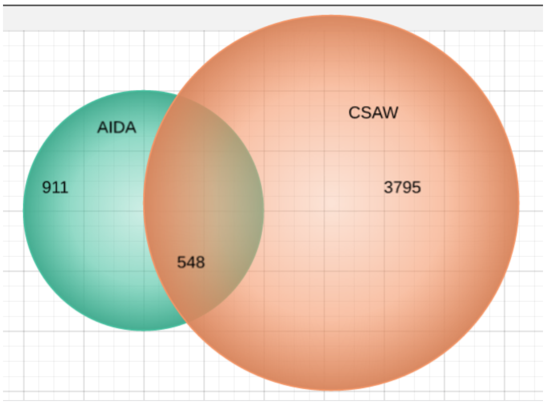

Figure 1: Entities Annotated by CSAW and AIDA

n	548
min	0.0333333333333333
max	10.0
$mean$	1.1219961591723266
std_dev	0.9419811831930233
$median$	1.0
$skewness$	5.1513317988869085
$kurtosis$	34.719172895503725

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

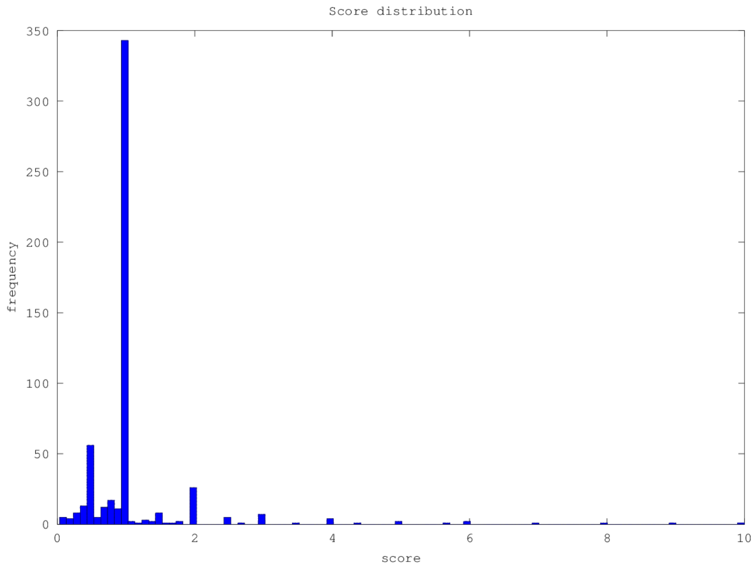

Figure 2: Distribution of scores for entities annotated by both CSAW and AIDA. CSAW and AIDA agree on 62% entities

Score	Frequency
0.03	1
0.17	1
0.20	3
0.25	7
0.31	1
0.33	7
0.50	52
0.60	1
0.67	9
1.00	342
1.50	8
2.00	26
4.00	4
4.33	1
5.00	2
5.67	1
6.00	2
7.00	1
8.00	1
9.00	1
10.00	1

Table 3: Distribution of $score = \frac{\text{Annotations by AIDA}}{\text{Annotations by CSAW}}$

2.2 Setup 2

AIDA was run in *CocktailPartyDisambiguation* mode. The Corpus used was yago trained from **2012** version of Wikipedia. A total of 104 documents were annotated.

Total entities annotated by AIDA	911
Total entities annotated by CSAW Team	3795
Total entities common to both	510
CSAW Entities missed by AIDA : (full list follows)	3285
AIDA Entities missed by CSAW : (full list follows)	401

Table 4: Statistics on Entities

n	510
min	0.045454545454545456
max	9.0
$mean$	1.1572691785583593
std_dev	0.9019959891067724
$median$	1.0
$skewness$	4.558145927208736
$kurtosis$	27.172766135955523

Table 5: Statistics for $score = \frac{\text{Annotations by AIDA}}{\text{Annotations by CSAW}}$

3 Analysis

- **AIDA annotates 989 entities, but only 548 of them were also in CSAW** There can be several reasons for this :
  - **Ambiguity of tagging**. Gandhi was wrongly linked to "Indira Gandhi" and not "Mahatma Gandhi" at several places.
  - **Different Wikipedia titles for same entity**. "Reuters" can be "Thomson Reuters" or "the news agency". Since the comparison was based on the title of the Wikipedia page that is linked, such annotations fail to match.
  - **New Wikipedia pages** might have emerged since CSAW was annotated. For eg, Harald zur Hansen is present in the corpus but not tagged by CSAW team.

- out of 548 matches, 342 (62%) were *exactly* annotated. This means that AIDA performs as good as a human being for these cases.

- Annotated data had some discrepancies. In file 13OctAmitSport14.txt, only 1 out of 9 instances of Steve Hodge was tagged. (Steve Hodge was mentioned by just the last name in 8 places)

- There were some extraneous tags. Words like Loss, Excellency, Utility, Year etc. were tagged with links to wiki pages. It is not exactly clear why such tags will be required. **The low recall**, (0.2606, $\frac{989}{3795}$) seems to be caused by such extraneous tags.

- There was not much difference in AIDA's performance with both local and global disambiguation settings.

- Resolving ambiguities is still a big challenge.