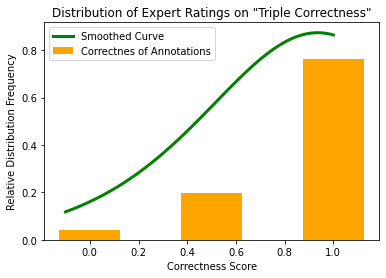
**Dataset Statistics**

* **Ground Truth or Verification Phase:**
  + Total Sentences: 80
  + Total Triples: 317
  + Verified: 325 = 102.52 %
  + Newly Added & Verified: 23 = 15/325 = 4.61%
  + New & Incorrect: 9
  + **Non-Sensical: 12 = 12/317 = 3.78%**
  + **Partially Incorrect & Updated: 63 = 63/317 = 19.87%**
  + **Fully Correct: 242 = 76.34%**
* **Distribution**

|  |  |  |
| --- | --- | --- |
|  | Variance |  |
| Annotators | 0 |  |
| Experts | 0.070863115 |  |
|  |  |  |
|  |  |  |
| z-Test: Two Sample for Means |  |  |
|  | *Annotators* | *Experts* |
| Mean | 1 | 0.861198738 |
| Known Variance | 1E-12 | 0.070863115 |
| Observations | 317 | 317 |
| Hypothesized Mean Difference | 0 |  |
| z | 9.283525311 |  |
| P(Z<=z) one-tail | 0 |  |
| z Critical one-tail | 1.644853627 |  |
| P(Z<=z) two-tail | 0 |  |
| z Critical two-tail | 1.959963985 |  |

****

A picture containing text, screenshot, plot, diagram

Description automatically generated

**EVALUATION**

* Implemented Tools/Libraries:
  1. SpaCy (SVO) Extraction Model
  2. Stanford OpenIE Triple Extractor
  3. GPT4 via ChatGPT Prompt
* Evaluation Criteria:
  1. Triple Mention Vectors
  2. Vector Similarity (Ground Truth VS Extracted Triple)
* Similarity Measures:
  1. Cosine Similarity
  2. Jaccard Similarity
* Evaluation Framework:
  1. Removing Similar Triples from Tools’ Output
  2. Calculating Each triple’s similarity score with GT Triples and keeping the highest similarity

**Pseudo Code:**

Triples (T)=[t1,t2,t3,…,tn], Ground\_Truth (GT)= [gt1,gt2,gt3,…,gtn]

Relatedness(T)=0

Uniform Case on all Text Triples

Average = MAX[LEN(T), LEN(GT)]

for ti in T:

for gtj in GT:

Similarity(tij) = **COSINE/JACCARD** (ti,gtj)

Similarity\_ti.append(Similarity(tij))

Score(ti)= MAX(Similarity\_ti[])

Relatedness(O) += Score(ti)

Relatedness(O) 🡨 Penalty|Amnesty (Based on Threshold)

**F-Measure (Average) = Relatedness(O)/ Average**

**F-Measure 🡨 Penalty|Amnesty (Based on Threshold)**

**THRESHOLDS**

|  |  |  |
| --- | --- | --- |
|  |  | Similarity |
| Amnesty (A) | Correct (1) | 0.81 – 1.00 |
|  | Partially Correct | 0.5 – 0.8 |
| Penalty (P) | Incorrect (0) | < 0.5 |

Table: Rescaling

**Challenges to Address**

* Filtering identical triples in from the model’s output i.e., OpenIE generates many identical triples.
* Refinement of Threshold Measure
* Similarity Measurement for triple or triple’s component-based?

**EVALUATION RESULTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cosine | Jaccard | Cosine | Jaccard |
| **Spacy** | 10.142 | 7.557 | 12.67% | 9.45% |
| **OpenIE** | 44.484 | 34.702 | 55.60% | 43.37% |
| **GPT 4** | **47.723** | **40.130** | **59.65%** | **50.162%** |

Table: Evaluation Results of Relatedness

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cosine | Jaccard | Cosine | Jaccard |
| **Spacy** | 10.452 | 7.614 | 13.06% | 9.52% |
| **OpenIE** | 48.601 | 35.856 | 60.75% | 44.82% |
| **GPT 4** | **51.924** | **42.386** | **64.91%** | **52.98%** |

Table: Evaluation Results with Threshold Amnesty

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cosine | Jaccard | Cosine | Jaccard |
| **Spacy** | 3.001 | 1.166 | 3.75% | 1.45% |
| **OpenIE** | 31.938 | 12.082 | 39.92% | 15.10% |
| **GPT 4** | **35.028** | **21.510** | **43.78%** | **26.88%** |

Table: Evaluation Results with Threshold Penalty

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cosine | Jaccard | Cosine | Jaccard |
| **Spacy** | 3.185 | 1.166 | 3.98% | 1.45% |
| **OpenIE** | 36.088 | 13.808 | 45.11% | 17.26% |
| **GPT 4** | **40.446** | **24.433** | **50.56%** | **30.54%** |

Table: Evaluation Results with Threshold (P & A)

