The Ground Truth Grind

Strategies for evaluating the quality of labeled data

Teresa O’Neill
Solutions Architect, Natural Language Services
● What else can we do besides stir the pot?

● Pour better data in!
  ○ Training data for supervised learning
  ○ Evaluation / ground truth datasets

● How can you trust your labeled data?

https://xkcd.com/1838/
Building Trust in Data Quality

Processes

Metrics

Transparency

Trust
We Are iMerit

iMerit leverages human intelligence to label and enrich data.

We power algorithms in machine learning and computer vision.

We effect **positive social and economic change**. We tap into a talent pool which was under-resourced and digitally excluded.

<table>
<thead>
<tr>
<th>2,700+</th>
<th>100+</th>
<th>9</th>
<th>&lt; 5%</th>
<th>24x7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employees</td>
<td>Clients</td>
<td>Centers</td>
<td>Attrition</td>
<td>Operations</td>
</tr>
</tbody>
</table>
iMerit Expertise

**Computer Vision**
- Polygon Annotation, LiDAR annotation, Bounding Boxes, Segmentation

**Natural Language Processing**
- Dialogue Annotation, NER, Intent Classification, Salience and Sentiment

**Content Services**
- Categorization, Mining, Moderation, Transcription, Product Enrichment

**Customer Support**
- Support desk, eCommerce Support, e-Learning Support, Onboarding, Verification
Transparency

- What does “high quality” mean?
- What assumptions do we start with?
- What expectations do we and our labelers have of each other?
What does “high quality” mean?

High-quality labeled data is...

- Complete
- Relevant
- Unbiased
- Consistent
- Accurate
What does “high quality” mean?

High-quality labeled data is... use-case dependent

- Complete
- Relevant
- Unbiased
- Consistent
- Accurate

what’s good for the goose… may not be good for the gander

Wikipedia commons
What assumptions do we start with?

<table>
<thead>
<tr>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality is measurable</td>
</tr>
<tr>
<td>Ground truth exists</td>
</tr>
<tr>
<td>Ground truth is knowable</td>
</tr>
<tr>
<td>Consistency indicates accuracy</td>
</tr>
<tr>
<td>Consistency is desirable</td>
</tr>
</tbody>
</table>
What assumptions do we start with?

- Quality is measurable
- Ground truth exists
- Ground truth is knowable
- Consistency indicates accuracy
- Consistency is desirable

“What do I know?”
What assumptions do we start with?

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality is measurable</td>
<td>… but not for all data points</td>
</tr>
<tr>
<td>Ground truth exists</td>
<td>… but not if the task is subjective or unbounded</td>
</tr>
<tr>
<td>Ground truth is knowable</td>
<td>… but not if the datapoint is ambiguous</td>
</tr>
<tr>
<td>Consistency indicates accuracy</td>
<td>… but not if it indicates bias</td>
</tr>
<tr>
<td>Consistency is desirable</td>
<td>… but not if it eliminates insights</td>
</tr>
</tbody>
</table>
What shared expectations do we have?

Success requires mutual investment in

- Guidelines
- Definitions of quality parameters
- Development of metrics appropriate for use case
- Feedback cycle
- Transparent reporting
Processes

- Guidelines & training
- Workflow customization
- Quality control
- Quality assurance

Trust

Transparency

Metrics
Quality Control vs. Quality Assurance

**QC**: Workflow designed to detect and correct defects

**QA**: Post-hoc audit designed to measure quality of a dataset
Quality Control

QC: Workflow designed to detect and correct defects

- Optimal method depends on use case
Quality Control

QC: Workflow designed to detect and correct defects

- Optimal methods depend on use case
- **Manual vs. automatic processes**
Quality Control

QC: Workflow designed to detect and correct defects

- Optimal methods depend on use case
- Manual vs. automatic processes
- **Process structure**
  - Multiple annotation
  - Multiple pass (expert reviewer)
  - Multiple annotation with adjudicator
Quality Control

QC: Workflow designed to detect and correct defects

- Optimal methods depend on use case
- Manual vs. automatic processes
- Process structure
- Interpretation of results

QC Feedback Cycle

- Iteration
- Insight
- Improvement
Quality Assurance

Owner: Client/internal team
   Expert labelers or domain experts
   Shared responsibility, higher cost burden on internal team during initial stage

Method:  Double-annotation w/ consistency measures
         Post-hoc comparison w/ random gold sample
         Benchmarking
Metrics: More of an art than a science

- Consistency vs. accuracy
- Priority alignment
- Transparent, actionable reports
The best things about NOLA are the waitstaff and the atmosphere. It's a lovely space, and whoever trains the staff is a pro: they are friendly and very efficient. The food is just average. I ordered a crab cake, it was virtually flavorless and the cornbread was dry as dust. A standout for us was dessert: delicious bread pudding and turtle pie. We will be coming back for the ambiance and deserts.
1 S: Good morning. Welcome to Made-Up Coffee Place. How can I help you?

2 C: Hi. Can I get two drinks, um, one extra hot large half decaf hazelnut latte and one cold brew, um, with the cold foam on that.

3 S: Sure, that's ...

4 C: Actually, on the first one, make that a triple shot.
Metrics: More of an art than a science

Set priorities

- Macro level vs. micro level
- Expected vs. actual distribution of classes
- Objective vs. subjective categories
- Major vs. minor error categories
- Primary vs. dependent labels

{We|PRON} attended {Conversational Interaction|CONFERENCE}...

{We|PRON} attended {Conversational Interaction|COMPANY}...

{We|PRON} attended Conversational Interaction...
Metrics: More of an art than a science

Set priorities

- Macro level vs. micro level
- Expected vs. actual distribution of classes
- Objective vs. subjective categories
- Major vs. minor error categories
- Primary vs. dependent labels

{We|PRON_k} attended {Conversational Interaction|CONFERENCE_j}... {it|PRON_j} was awesome

{We|PRON_k} attended {Conversational Interaction|COMPANY_j}... {it|PRON_j} was awesome
Metrics: More of an art than a science

Set priorities

Build transparent, actionable reports
  • Insights drive improvement

<table>
<thead>
<tr>
<th>Entity Label Error Categories</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect matches</td>
<td>90%</td>
</tr>
<tr>
<td>Misses</td>
<td>2%</td>
</tr>
<tr>
<td>Additions</td>
<td>3%</td>
</tr>
<tr>
<td>Pure label errors--diff maj. class</td>
<td>1%</td>
</tr>
<tr>
<td>Pure label errors--diff min. class</td>
<td>2%</td>
</tr>
<tr>
<td>Pure span error--nesting</td>
<td>0.2%</td>
</tr>
<tr>
<td>Pure span error--shifting</td>
<td>0.4%</td>
</tr>
<tr>
<td>Label+span error...</td>
<td>...</td>
</tr>
</tbody>
</table>
Building Trust in Data Quality
Thank You!

teresa@imerit.net