Got Data?
The Importance of High-Quality Training Data for Building Effective Language-Based Solutions
Appen: Data with a Human Touch

Experience working in 130+ countries

Expertise in 180+ languages
Get the full list →

20+ years working with leading global technology companies

Access to a curated crowd of over 400,000 flexible workers worldwide

Over 1 billion judgments made and 500,000 hours of audio processed

Over 330 employees located in six offices around the globe
The need for annotated data

The role of human-annotated linguistic data in development of AI systems that include NLP components is critical but often unacknowledged.

Where does the data come from?
(graduate students?)
Choice 1: Buy it or Build it?

Licensable Public Data?

Pros:
- Cheap!
- Results easily compared
- Known quality thresholds

Cons:
- Not specific to your domain
- Solves somebody else’s problem
- Language coverage

Roll your own?

Pros:
- There’s no data like your own data
- Tailor labels to your problem
- In your target (language) market

Cons:
- Must define the label set
- Must find and train annotators
- Quality measures?
Choice 2: What’s my label set?

Rich Linguistic Features

Crowdsourced labeling

Table 2: The Penn Treebank POS tagset.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordinating conjunction</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>EX</td>
<td>Existential there</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition/subordinating conjunction</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>JJR</td>
<td>Adjective, comparative</td>
</tr>
<tr>
<td>JJS</td>
<td>Adjective, superlative</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, singular or mass</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
</tr>
<tr>
<td>PRP</td>
<td>Personal pronoun</td>
</tr>
<tr>
<td>PP$</td>
<td>Possessive pronoun</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
</tr>
<tr>
<td>SYM</td>
<td>Symbol (mathematical or scientific)</td>
</tr>
</tbody>
</table>

2018-02-07
Sentiment Analysis: Movie Review Domain

The Last Jedi is a bomb!
The Iron Triangle ...

Quality

Time

Cost

Pick 2
Choice 3: Complexity or Volume?

Complexity of Task

Time

Volume of Data
Data Annotation LOE (Sentiment)

- Documents: Tweets or tweet-like social media text (12-200 chars)
- 3-way judgment at document level (Positive/Negative/Neutral)
- Minimal training requirements (1-2 page instructions)
- Throughput: 250 documents per hour
  - A bit slower depending on the writing system, e.g. 180 per hour for Chinese

→ 10K docs = 40 hours of labor for 1 judgment
How useful is a single judgment?

Pretty useless, actually ...

- 30-40% disagreement rate on this task for two minimally trained judges
- 90% consensus with screening and adding a third judge (2 or more judges agree)
Making choices …

- No magic bullets for annotated data creation, only tradeoffs
- Budgeting for data creation is indispensable
- Better results with data that is specific to your domain, annotated to your specifications
  - (But be pragmatic about quality targets for newly-designed labeling tasks)
- Small trained/screened crowds with quality feedback loop can yield good results on modest volumes of data
Thank you

James Lyle
Director Custom Linguistic Solutions
jlyle@appen.com