Multi-Modal Information Extraction (Question-Answer Framework)
Vamsi Chitters, Fabian Frank, Lars Jebe
{vamsikc, fabfrank, larsjebe}@stanford.edu

Human reliance on various modes (e.g., visual, auditory, textual) to understand context is often achieved through supervised learning. Multi-modal information extraction similarly aspires to extract domain knowledge using various modes.

**Potential applications:**
- Answer questions such as “what is the best car in the city?” (using reviews and images)
- Structure info (e.g., Knowledge Graph), NER, Q&A

**Evaluation**
- Data quality really matters (e.g., web scraped data)
- Domain knowledge (manually selected features) improves accuracy

**Baseline:**
- self-trained Word2Vec with single-word value predictions using cosine similarity

**LSTM-Pointer Network:**
- LSTM layer for description, custom embedding layer for attributes, model outputs value start and end positions within description
- LSTM represents text as context vector
- Attribute embedding captures query intention
- Predicts start and end of value independently

**Recommendation-based Pointer Network:**
- Constructs recommendations (hand-crafted features) based on prior distr. of values, captures dependence of value start and end positions
- Predicts span instead of independent start and end positions
- Additional hidden layer to learn weighting recommendations

**DISCUSSION**
- Data quality really matters (e.g., web scraped data contains uninformative ground truth values)
- Difficult to replicate human-like multi-modal information extraction
- Expected: using two modes would improve task accuracy
- Reality: images don’t help much (hard to extract semantic meaning)
- LSTM models proved effective for QA due to context (well-researched)
- Image captions predicts classes (nouns), not fine-grained predictions (adjectives)
- Domain knowledge (manually selected features) improves accuracy
- Pointer network addresses multi-word value prediction effectively - captures dependency between start and end index well

**FUTURE WORK**
- Gain intuition from more complex models (e.g. R-NET)
- Explore other multi-word value approaches (e.g. BIO-Tagging)
- Enhance attention model to focus on relevant text and image parts (i.e., look at car body if query is “car brand”)

**RESULTS AND ANALYSIS**

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Attribute</th>
<th>Prediction</th>
<th>Actual Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Word (close)</td>
<td>batteries included</td>
<td>included</td>
<td>yes</td>
</tr>
<tr>
<td>Multi-Word (close)</td>
<td>measurements</td>
<td>height 1.17 m</td>
<td>height 1.17 m</td>
</tr>
<tr>
<td>Complete Miss</td>
<td>finish</td>
<td>price</td>
<td>semi-gloss</td>
</tr>
</tbody>
</table>

1 Other errors included predicting e.g., $v_{3 \rightarrow 5}$ instead of $v_{3 \rightarrow 4}$ (SE-Tagging). Improvements for all error types include: (1) pre-train value-categories for attributes (2) impose mutual dependence between start and end position (3) add domain-aware constraints such as (value for “dimensions” must include numbers)