OpenTag: Open Attribute Value Extraction from Product Profiles

Guineng Zheng*, Subhabrata MukherjeeA, Xin Luna DongA, FeiFei Li*

AAmazon.com, *University of Utah
{subhomj, lunadong}@amazon.com, {guineng, lifeifei}@cs.utah.edu

KDD 2018

Amazon Catalog provides structured information on product attributes. This information is often noisy or missing for a lot of attributes

We develop OpenTag: a novel deep tagging model to discover missing values of attributes from unstructured product profiles like title, description, and bullets

OpenTag does not rely on dictionaries of values or hand-crafted features

It discovers new attribute values never encountered before with an open world assumption

It uses active learning to reduce manual annotation effort by 3.3x

OpenTag Components: Extraction via Sequence Tagging

- Word Embeddings model distributed attribute-value representations
- Bidirectional LSTM (BiLSTM) captures contextual information from long and short range dependencies via hidden states: $h_i = σ(\tilde{h}_i)$
- BiLSTM captures sequential nature of tokens but not tags
- Conditional Random Field (CRF) considers sequential nature of tags to extract coherent attribute values:
  $$\Pr(y|x; \Psi) = \prod_{t=1}^{T} \exp \left( \sum_{j=1}^{V} \phi_j f_k(y_{t-1}, y_t, \tilde{h}_t) \right)$$
- Attention mechanism captures importance of tags for tokens, important concepts, and generates interpretable explanation for its verdict: $I_t = \sum_{t'=1}^{V} r_{t',t} \cdot \tilde{h}_t$

OpenTag: Active Learning

- Uses active learning to reduce manual annotation effort
- To identify difficult-to-extract instances for the model, and ask for annotation
- Uses tag flips as a Query Strategy to identify difficult training instances
- Tag Flip: Change in tag of a token across successive epochs during training
- Frequent tag flips indicate OpenTag is uncertain about the sample; not stable

OpenTag: Architecture

- Uses active learning to reduce manual annotation effort
- To identify difficult-to-extract instances for the model, and ask for annotation
- Uses tag flips as a Query Strategy to identify difficult training instances
- Tag Flip: Change in tag of a token across successive epochs during training
- Frequent tag flips indicate OpenTag is uncertain about the sample; not stable

Extracting Structured Values from Unstructured Text

- Given a product title "Variety Pack Fillet Mignon and Porterhouse Steak Dog Food (12 Count)”: OpenTag extracts attribute values “size = 12 count” and “flavor = (Porterhouse Steak, Fillet Mignon)”
- It extracts structured multi-word values, and multiple values for an attribute
- Sequence Tagging: It exploits distributional semantics to assign tags to tokens to extract attribute values

### Sequence Tagging

<table>
<thead>
<tr>
<th>Sequence</th>
<th>duck</th>
<th>fillet</th>
<th>mignon</th>
<th>ranch</th>
<th>raised</th>
<th>lamb</th>
<th>flavor</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIO</td>
<td>O</td>
<td>B</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>E</td>
<td>O</td>
</tr>
<tr>
<td>BIO</td>
<td>O</td>
<td>B</td>
<td>E</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>E</td>
</tr>
<tr>
<td>IOB</td>
<td>B</td>
<td>O</td>
<td>B</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>I</td>
</tr>
</tbody>
</table>

OpenTag: Active Learning

- Uses active learning to reduce manual annotation effort
- To identify difficult-to-extract instances for the model, and ask for annotation
- Uses tag flips as a Query Strategy to identify difficult training instances
- Tag Flip: Change in tag of a token across successive epochs during training
- Frequent tag flips indicate OpenTag is uncertain about the sample; not stable

- Amazon Catalog provides structured information on product attributes. This information is often noisy or missing for a lot of attributes
- We develop OpenTag: a novel deep tagging model to discover missing values of attributes from unstructured product profiles like title, description, and bullets
- OpenTag does not rely on dictionaries of values or hand-crafted features
- It discovers new attribute values never encountered before with an open world assumption
- It uses active learning to reduce manual annotation effort by 3.3x

OpenTag Components: Extraction via Sequence Tagging

- Word Embeddings model distributed attribute-value representations
- Bidirectional LSTM (BiLSTM) captures contextual information from long and short range dependencies via hidden states: $h_i = σ(\tilde{h}_i)$
- BiLSTM captures sequential nature of tokens but not tags
- Conditional Random Field (CRF) considers sequential nature of tags to extract coherent attribute values:
  $$\Pr(y|x; \Psi) = \prod_{t=1}^{T} \exp \left( \sum_{j=1}^{V} \phi_j f_k(y_{t-1}, y_t, \tilde{h}_t) \right)$$
- Attention mechanism captures importance of tags for tokens, important concepts, and generates interpretable explanation for its verdict: $I_t = \sum_{t'=1}^{V} r_{t',t} \cdot \tilde{h}_t$