OpenTag: Open Attribute Value Extraction From Product Profiles

Guineng Zheng*, Subhabrata MukherjeeΔ, Xin Luna DongΔ, FeiFei Li*
ΔAmazon.com, *University of Utah
Guineng Zheng
Motivation

Alexa, what are the flavors of Nescafe?

Nescafe Coffee flavors include caramel, mocha, vanilla, coconut, cappuccino, original/regular, decaf, espresso, and cafe au lait.
Problem Statement: Extract attribute values from (text of) product profiles

<table>
<thead>
<tr>
<th><strong>Input</strong> Product Profile</th>
<th><strong>Output</strong> Extractions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>1. filet mignon 2. porterhouse steak</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>CESAR Canine Cuisine Variety Pack Filet Mignon &amp; Porterhouse Steak Dog Food (Two 12-Count Cases) A Delectable Meaty Meal for a Small Canine Looking for the right food... This delicious dog treat contains tender slices of meat in gravy and is formulated to meet the nutritional needs of small dogs.</td>
</tr>
<tr>
<td><strong>Bullets</strong></td>
<td>• Filet Mignon Flavor; • Porterhouse Steak Flavor; • CESAR Canine Cuisine provides complete and balanced nutrition...</td>
</tr>
<tr>
<td><strong>Flavor</strong></td>
<td>cesar</td>
</tr>
<tr>
<td><strong>Brand</strong></td>
<td>canine cuisine</td>
</tr>
</tbody>
</table>

...
Characteristics of Attribute Extraction

**Limited semantics, irregular syntax**
- Most titles have **10-15** words
- Most bullets have **5-6** words
- **Phrases** not Sentences
  - Lack of regular grammatical structure in titles and bullets
  - Attribute stacking

**Open World Assumption**
- No Predefined Attribute Value
- New Attribute Value Discovery

1. Rachael Ray Nutrish Just 6 Natural Dry Dog Food, Lamb Meal & Brown Rice Recipe
2. Lamb Meal is the #1 Ingredient

1. beef flavor
2. lamb flavor
3. meat in gravy flavor
Contributions and Prior Work (to do)
Outline

• Sequence Tagging
• Models
• Active Learning
• Experiments and Discussions
Attribute Extraction as Sequence Tagging

- **B**: Beginning of attribute value
- **I**: Inside of attribute value
- **O**: Outside of attribute value
- **E**: End of attribute value

\[ x = \{ w_1, w_2, \ldots, w_n \} \text{ input sequence} \]

\[ y = \{ t_1, t_2, \ldots, t_n \} \text{ tagging decision} \]

Flavor Extractions:

\{beef meal\} → \{ranch raised lamb\}
Models

- BiLSTM
- BiLSTM + CRF
- Attention Mechanism
- OpenTag Architecture
OpenTag Architecture

CRF Layer

Attention Mechanism

BiLSTM Layer

Word Embedding

KDD 2018
Word Embedding

• Map words co-occurring in a similar context to nearby points in embedding space

• Pre-trained embeddings learn single representation for each word
  • But ‘duck’ as a Flavor should have different embedding than ‘duck’ as a Brand

• OpenTag learns word embeddings conditioned on attribute-tags
Bi-directional LSTM

- LSTM (Hochreiter, 1997) capture long and short range dependencies between tokens, suitable for modeling token sequences.

- Bi-directional LSTM’s improve over LSTM’s capturing both forward ($f_t$) and backward ($b_t$) states at each timestep ‘t’.

- Hidden state $h_t$ at each timestep generated as: $h_t = \sigma([b_t, f_t])$
Conditional Random Fields (CRF)

- Bi-LSTM captures dependency between token sequences, but not between output tags
- Likelihood of a token-tag being ‘E’ (end) or ‘I’ (intermediate) increases, if the previous token-tag was ‘I’ (intermediate)
- Given an input sequence $x = \{x_1, x_2, ..., x_n\}$ with tags $y = \{y_1, y_2, ..., y_n\}$: linear-chain CRF models:

$$\Pr(y|x; \Psi) \propto \prod_{t=1}^{T} \exp \left( \sum_{k=1}^{K} \psi_k f_k(y_{t-1}, y_t, x) \right)$$
Bi-directional LSTM + CRF

$$\Pr(y|x; \Psi) \propto \prod_{t=1}^{T} \exp \left( \sum_{k=1}^{K} \psi_k f_k(y_{t-1}, y_t, \langle h_t \rangle) \right)$$

CRF feature space formed by Bi-LSTM hidden states
Attention Mechanism

• Not all hidden states equally important for the CRF
• Focus on important concepts, downweight the rest => attention!
• Attention matrix $A$ to attend to important BiLSTM hidden states ($h_t$)
  • $\alpha_{t,t'} \in A$ captures similarity between $h_t$ and $h_{t'}$
• Attention-focused representation $l_t$ of token $x_t$ given by:

$$l_t = \sum_{t'=1}^{n} \alpha_{t,t'} \cdot h_{t'}$$
OpenTag Architecture

CRF Layer

Attention Mechanism

BiLSTM Layer

Word Embedding

KDD 2018
Final Classification

Maximize log-likelihood of joint distribution

\[
L(\Psi) = \sum_{i=1}^{m} \log \Pr(y_i|x_i; \Psi)
\]

Best possible tag sequence with highest conditional probability

\[
y^* = \arg\max_y \Pr(y|x; \Psi)
\]
## Experimental Discussions: Datasets

<table>
<thead>
<tr>
<th>Domain</th>
<th>Profile</th>
<th>Attribute</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Samples</td>
<td>Extractions</td>
</tr>
<tr>
<td>Dog Food (DS)</td>
<td>Title</td>
<td>Flavor</td>
<td>470</td>
<td>876</td>
</tr>
<tr>
<td>Dog Food</td>
<td>Title</td>
<td>Flavor</td>
<td>470</td>
<td>716</td>
</tr>
<tr>
<td></td>
<td>Desc</td>
<td>Flavor</td>
<td>450</td>
<td>569</td>
</tr>
<tr>
<td></td>
<td>Bullet</td>
<td>Flavor</td>
<td>800</td>
<td>1481</td>
</tr>
<tr>
<td></td>
<td>Title</td>
<td>Brand</td>
<td>470</td>
<td>480</td>
</tr>
<tr>
<td></td>
<td>Title</td>
<td>Capacity</td>
<td>470</td>
<td>428</td>
</tr>
<tr>
<td></td>
<td>Title</td>
<td>Multi</td>
<td>470</td>
<td>1775</td>
</tr>
<tr>
<td>Camera</td>
<td>Title</td>
<td>Brand</td>
<td>210</td>
<td>210</td>
</tr>
<tr>
<td>Detergent</td>
<td>Title</td>
<td>Scent</td>
<td>500</td>
<td>487</td>
</tr>
<tr>
<td>Datasets/Attribute</td>
<td>Models</td>
<td>Precision</td>
<td>Recall</td>
<td>Fscore</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------</td>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Dog Food: Title</td>
<td>BiLSTM</td>
<td>83.5</td>
<td>85.4</td>
<td>84.5</td>
</tr>
<tr>
<td>Attribute: Flavor</td>
<td>BiLSTM-CRF</td>
<td>83.8</td>
<td>85.0</td>
<td>84.4</td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>86.6</td>
<td>85.9</td>
<td>86.3</td>
</tr>
<tr>
<td>Camera: Title</td>
<td>BiLSTM</td>
<td>94.7</td>
<td>88.8</td>
<td>91.8</td>
</tr>
<tr>
<td>Attribute: Brand name</td>
<td>BiLSTM-CRF</td>
<td>91.9</td>
<td>93.8</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>94.9</td>
<td>93.4</td>
<td>94.1</td>
</tr>
<tr>
<td>Detergent: Title</td>
<td>BiLSTM</td>
<td>81.3</td>
<td>82.2</td>
<td>81.7</td>
</tr>
<tr>
<td>Attribute: Scent</td>
<td>BiLSTM-CRF</td>
<td>85.1</td>
<td>82.6</td>
<td>83.8</td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>84.5</td>
<td>88.2</td>
<td>86.4</td>
</tr>
<tr>
<td>Dog Food: Description</td>
<td>BiLSTM</td>
<td>57.3</td>
<td>58.6</td>
<td>58</td>
</tr>
<tr>
<td>Attribute: Flavor</td>
<td>BiLSTM-CRF</td>
<td>62.4</td>
<td>51.5</td>
<td>56.9</td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>64.2</td>
<td>60.2</td>
<td>62.2</td>
</tr>
<tr>
<td>Dog Food: Bullet</td>
<td>BiLSTM</td>
<td>93.2</td>
<td>94.2</td>
<td>93.7</td>
</tr>
<tr>
<td>Attribute: Flavor</td>
<td>BiLSTM-CRF</td>
<td>94.3</td>
<td>94.6</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>95.7</td>
<td>95.7</td>
<td>95.7</td>
</tr>
<tr>
<td>Dog Food: Title</td>
<td>BiLSTM</td>
<td>71.2</td>
<td>67.4</td>
<td>69.3</td>
</tr>
<tr>
<td>Multi Attribute:</td>
<td>BiLSTM-CRF</td>
<td>72.9</td>
<td>67.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Brand, Flavor, Capacity</td>
<td>OpenTag</td>
<td>76.0</td>
<td>68.1</td>
<td>72.1</td>
</tr>
</tbody>
</table>
Discovering new attribute-values not seen during training

<table>
<thead>
<tr>
<th>Train-Test Framework</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disjoint Split (DS)</td>
<td>83.6</td>
<td>81.2</td>
<td>82.4</td>
</tr>
<tr>
<td>Random Split</td>
<td>86.6</td>
<td>85.9</td>
<td>86.3</td>
</tr>
</tbody>
</table>
Interpretability via Attention
OpenTag achieves better concept clustering

Distribution of word vectors before attention

Distribution of word vectors after attention
Semantically related words come closer in the embedding space
Active Learning (Settles, 2009)

- Query selection strategy like *uncertainty sampling* selects sample with *highest uncertainty* for annotation
- Ignores difficulty in estimating *individual tags*
Tag Flip as Query Strategy

- Simulate a committee of OpenTag learners $C$ over epochs
- Most informative sample => major disagreement among committee members for tags of its tokens
- Use dropout mechanism for simulating committee of learners

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

Tag flips = 4

- Most informative sample has *highest tag flips* across all the epochs
OpenTag reduces burden of human annotation by 3.3x

Learning from scratch on detergent data

- 150 labeled samples
  - Precision: 71.8%
  - Recall: 63.8%

- 500 labeled samples
  - Precision: 90.7%
  - Recall: 92.4%
  - 94.2%

Learning from scratch on multi extraction

- 50 labeled samples
  - Precision: 81.9%
  - Recall: 60.4%

- 150 labeled samples
  - Precision: 76.4%
  - Recall: 73.6%
  - 78.5%

KDD 2018
## Production Impact

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Increase in Coverage over Existing Production System (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute_1</td>
<td>53</td>
</tr>
<tr>
<td>Attribute_2</td>
<td>45</td>
</tr>
<tr>
<td>Attribute_3</td>
<td>50</td>
</tr>
<tr>
<td>Attribute_4</td>
<td>48</td>
</tr>
</tbody>
</table>
Summary

• OpenTag model based on word embeddings, Bi-LSTM, CRF and attention
  • Open world assumption (OWA), multi-word and multiple attribute value extraction
• OpenTag + Active learning reduces burden of human annotation (by 3.3x)
  • Method of tag flip as query strategy
• Interpretability
  • Better concept clustering, interpretability via attention, etc.
Backup Slides
Multiple attribute values

• Predicting multiple attribute values **jointly**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand: Single</td>
<td>52.6</td>
<td>42.6</td>
<td>47.1</td>
</tr>
<tr>
<td>Brand: Multi</td>
<td>58.4</td>
<td>44.7</td>
<td>50.6</td>
</tr>
<tr>
<td>Flavor: Single</td>
<td>83.6</td>
<td>81.2</td>
<td>82.4</td>
</tr>
<tr>
<td>Flavor: Multi</td>
<td>83.7</td>
<td>77.5</td>
<td>80.5</td>
</tr>
<tr>
<td>Capacity: Single</td>
<td>81.5</td>
<td>86.4</td>
<td>83.9</td>
</tr>
<tr>
<td>Capacity: Multi</td>
<td>87.0</td>
<td>87.2</td>
<td>87.1</td>
</tr>
</tbody>
</table>

• Modify tagging strategy to have separate tag-set \{B_a, I_a, O_a, E_a\} for each attribute ‘a’
Why Sequence Tagging

Open World Assumption & Label Scaling
- Limited Tags: [BIOE]
- Unlimited Attributes
  - Tag-set not attribute-specific

<table>
<thead>
<tr>
<th>B</th>
<th>E</th>
<th>Detected Flavors</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>B</td>
<td>E</td>
</tr>
<tr>
<td>lamb</td>
<td>flavor</td>
<td>Australian lamb</td>
</tr>
<tr>
<td>beef</td>
<td>and</td>
<td>green lentils</td>
</tr>
</tbody>
</table>

Discovering multi-word & multiple attribute values
- Semantics of word itself and surrounding context for chunking

<table>
<thead>
<tr>
<th>Tag</th>
<th>Evidence of Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>dry dog food, duck, 10lb</td>
<td>duck itself</td>
</tr>
<tr>
<td>whitefish flavor</td>
<td>keyword flavor</td>
</tr>
<tr>
<td>lamb recipe</td>
<td>lamb, keyword recipe</td>
</tr>
<tr>
<td>beef and green lentils</td>
<td>beef, conjunct word “and”</td>
</tr>
</tbody>
</table>
Bi-directional LSTM

\[ \Pr(y_t = k) = \text{softmax}(h_t \cdot W_h) \]

**Word Index**

**Word Embedding**
- glove embedding 50

**Hidden Vector**
- 100+100=200 units

**Backward LSTM**
- 100 units

**Forward LSTM**
- 100 units

**Cross Entropy Loss**

- ranch raised beef flavor
- w1 w2 w3 w4
- e1 e2 e3 e4
- b1 b2 b3 b4
- f1 f2 f3 f4
- h1 h2 h3 h4
Bi-directional LSTM + CRF

$$\Pr(y|x; \Psi) \propto \prod_{t=1}^{T} \exp \left( \sum_{k=1}^{K} \psi_{k} f_{k}(y_{t-1}, y_{t}, \langle h_{t} \rangle) \right)$$

Cross Entropy Loss

Conditional Random Field

CRF feature space formed by Bi-LSTM hidden states

Forward LSTM
100 units

Embedding
glove embedding 50

Word Index

KDD: ranch
ranch

raised

beef

flavor
Uncertainty Sampling: Probability as Query Strategy

• Select instance with maximum uncertainty
  • Best possible tag sequence from CRF:
    \[ y^* = \arg\max_y \Pr(y|x; \Psi) \]
  • Label instance with maximum uncertainty:
    \[ Q^{lc}(x) = 1 - \Pr(y^*|x; \Psi) \]

• Considers entire label sequence \( y \), ignores difficulty in estimating individual tags \( y_t \in y \)
Tag Flip as Query Strategy

- Most informative instance has maximum tag flips aggregated over all of its tokens across all the epochs:

\[ Q^{tf}(x) = \sum_{e=1}^{E} \sum_{t=1}^{n} I(y^*_t(\Psi^{(e-1)}) \neq y^*_t(\Psi^{(e)})) \]

- Top \( B \) samples with the highest number of flips are manually annotated with tags.
Experiments and Discussions
Active Learning: Tag Flip better than Uncertainty Sampling

TF v.v. LC on detergent data

TF v.v. LC on multi extraction