OpenTag: Open Attribute Value Extraction From Product Profiles

Guineng Zheng*, Subhabrata MukherjeeΔ, Xin Luna DongΔ, FeiFei Li*

ΔAmazon.com, *University of Utah
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.
Attribute value extraction from product profiles

Flavor

Variety Pack Filet Mignon and Porterhouse Steak Dog Food (12 Count)
Price: $92.60 & FREE Shipping
Be the first to review this item

- 6 trays of Filet Mignon flavor in meaty juices
- Cesar pet food has an irresistible taste with exceptional palatability to tempt even the fussiest dogs
- Formulated to meet the nutritional levels established by the AAFCO Dog Food Nutrient Profiles for maintenance

Brand

In stock.
Get it as soon as Wednesday, Feb. 14 when you choose Two-Day Shipping at checkout.
Ships from and sold by Cunningham Collective.

Product description
Variety pack includes: 6 trays of Filet Mignon flavor in meaty juices 6 trays of Porterhouse Steak flavor in meaty juices Cesar pet food has an irresistible taste with exceptional palatability to tempt even the fussiest dogs Formulated to meet the nutritional levels established by the AAFCO Dog Food Nutrient Profiles for maintenance Complete & balanced nutrition for small adult dogs Fortified with vitamins and minerals Packaged in convenient feeding trays with no-fuss, peel-away freshness seals Includes 6 Each Chicken & Liver
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. beef flavor
2. lamb flavor
3. venison flavor

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

<table>
<thead>
<tr>
<th></th>
<th>Open World Assumption</th>
<th>No Lexicon, No Hand-crafted Features</th>
<th>Active Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghani et al. 2003,</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Putthividhya et al.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang et al. 2015,</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Kozareva et al. 2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>OpenTag (this work)</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Outline

• Problem Definition
• Models
  • Experiments
• Active Learning
  • Experiments
Recap: Problem Statement

Given product profiles (e.g., titles, descriptions, bullets) and a set of attributes: extract values of attributes from profile texts

<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><strong>Description</strong></td>
</tr>
</tbody>
</table>
| CESAR Canine Cuisine Variety Pack Filet Mignon & 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. | • Filet Mignon Flavor;  
• Porterhouse Steak Flavor;  
• CESAR Canine Cuisine provides complete and balanced nutrition ... | 1.filet mignon 2.porterhouse steak | cesar canine cuisine |

KDD 2018
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,...,w_n\}** input sequence

**y=\{t_1,t_2,...,t_n\}** tagging decision

**INPUT**

- beef
- meal
- &
- ranch
- raised
- lamb
- recipe

**OUTPUT**

- {beef meal}
- {ranch raise lamb}

**Flavor Extractions**

- **t_1**
- **t_2**
- **t_3**
- **t_4**
- **t_5**
- **t_6**
- **t_7**

**x**

**y**
Outline

• Introduction

• Models
  • BiLSTM
  • BiLSTM + CRF
  • Attention Mechanism
  • OpenTag Architecture

• Active Learning
OpenTag Architecture

CRF Layer

Attention Mechanism

BiLSTM Layer

Word Embedding

KDD 2018
OpenTag Architecture (1/4): Word Embedding

Map ‘beef’, ‘chicken’, ‘pork’ to nearby points in Flavor–embedding space
Capture long and short range dependencies in input sequence via forward and backward hidden states
• Bi-LSTM captures dependency between token sequences, but not between output tags
• Conditional Random Field (CRF) enforces *tagging consistency*
Focus on important hidden concepts, downweight the rest => attention!

- Attention matrix $A$ to attend to important BiLSTM hidden states ($h_t$)
  - $\alpha_{t,t'} \in A$ captures importance of $h_t$ w.r.t. $h_{t'}$
  - 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
# 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><strong>86.6</strong></td>
<td><strong>85.9</strong></td>
<td><strong>86.3</strong></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><strong>93.8</strong></td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td><strong>94.9</strong></td>
<td>93.4</td>
<td><strong>94.1</strong></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></td>
<td>BiLSTM-CRF</td>
<td><strong>85.1</strong></td>
<td>82.6</td>
<td>83.8</td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>84.5</td>
<td><strong>88.2</strong></td>
<td><strong>86.4</strong></td>
</tr>
<tr>
<td>Dog Food: Description</td>
<td>BiLSTM</td>
<td>57.3</td>
<td>58.6</td>
<td>58.0</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><strong>64.2</strong></td>
<td><strong>60.2</strong></td>
<td><strong>62.2</strong></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><strong>95.7</strong></td>
<td><strong>95.7</strong></td>
<td><strong>95.7</strong></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><strong>76.0</strong></td>
<td><strong>68.1</strong></td>
<td><strong>72.1</strong></td>
</tr>
</tbody>
</table>

Overall, OpenTag obtains high F-score of 82.8%
### Results

<table>
<thead>
<tr>
<th>Datasets/Attribute</th>
<th>Models</th>
<th>Precision</th>
<th>Recall</th>
<th>Fscore</th>
</tr>
</thead>
<tbody>
<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></td>
<td><strong>85.9</strong></td>
<td><strong>86.3</strong></td>
<td></td>
</tr>
<tr>
<td>Dog Food: Description</td>
<td>BiLSTM</td>
<td>88.8</td>
<td>91.8</td>
<td></td>
</tr>
<tr>
<td>Attribute: Flavor</td>
<td>BiLSTM-CRF</td>
<td>93.8</td>
<td>92.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>93.4</td>
<td>94.1</td>
<td></td>
</tr>
<tr>
<td>Dog Food: Bullet</td>
<td>BiLSTM</td>
<td>82.2</td>
<td>81.7</td>
<td></td>
</tr>
<tr>
<td>Attribute: Flavor</td>
<td>BiLSTM-CRF</td>
<td>82.6</td>
<td>83.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OpenTag</td>
<td>84.3</td>
<td>86.4</td>
<td></td>
</tr>
<tr>
<td>Dog Food: Title</td>
<td>BiLSTM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi Attribute:</td>
<td>BiLSTM-CRF</td>
<td></td>
<td></td>
<td></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>

- Highest improvement in F-score of 5.3% over BiLSTM-CRF for product descriptions
- However, less accurate than titles
OpenTag discovers new attribute-values not seen during training with 82.4% F-score

<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>71.9</td>
<td>86.3</td>
</tr>
</tbody>
</table>

No overlap in attribute value between train and test splits
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
Outline

• Introduction

• Models
  • BiLSTM
  • BiLSTM + CRF
  • Attention Mechanism
  • OpenTag Architecture

• Active Learning
Active Learning: Motivation

• Annotating training data is expensive and time-consuming
  • Does not scale to thousands of verticals with hundreds of attributes and thousands of values in each domain
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 over multiple epochs
- Most informative sample => major disagreement among committee members for tags of its tokens across epochs
- 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
Tag Flip (red) better than Uncertainty Sampling (blue)

TF v.v. LC on detergent data

TF v.v. LC on multi extraction
OpenTag reduces burden of human annotation by 3.3x

- OpenTag requires only 500 training samples to obtain > 90% P-R
- Active learning brings it down to 150 training samples to match similar performance
## Production Impact

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Previous Coverage (%)</th>
<th>OpenTag Coverage (%)</th>
<th>Increase in Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute_1</td>
<td>23</td>
<td>78</td>
<td>53</td>
</tr>
<tr>
<td>Attribute_2</td>
<td>21</td>
<td>72</td>
<td>45</td>
</tr>
<tr>
<td>Attribute_3</td>
<td>&lt; 1</td>
<td>56</td>
<td>50</td>
</tr>
<tr>
<td>Attribute_4</td>
<td>&lt; 1</td>
<td>49</td>
<td>48</td>
</tr>
</tbody>
</table>
Summary

• OpenTag models open world assumption (OWA), multi-word and multiple attribute value extraction with sequence tagging
  • Word embeddings + Bi-LSTM + CRF + attention

• OpenTag + Active learning reduces burden of human annotation (by 3.3x)
  • Method of tag flip as query strategy

• Interpretability
  • Better concept clustering, attention heatmap, etc.
Summary

• OpenTag models open world assumption (OWA), multi-word and multiple attribute value extraction with sequence tagging
  • Word embeddings + Bi-LSTM + CRF + attention

• OpenTag + Active learning reduces burden of human annotation (by 3.3x)
  • Method of tag flip as query strategy

• Interpretability
  • Better concept clustering, attention heatmap, etc.

KDD 2018
Backup Slides
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: $h_t = \sigma([b_t, f_t])$
Bi-directional LSTM

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

- Cross Entropy Loss
- Hidden Vector
  \[ 100 + 100 = 200 \text{ units} \]
- Backward LSTM
  \[ 100 \text{ units} \]
- Forward LSTM
  \[ 100 \text{ units} \]
- Word Embedding
  \[ \text{glove embedding 50} \]
- Word Index

Word Index

- ranch
- raised
- beef
- flavor

- glove embedding 50
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

Cross Entropy Loss

CRF feature space formed by Bi-LSTM hidden states

Forward LSTM
100 units

Embedding
glove embedding 50

Word Index
KDD: ranch raised beef flavor

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)
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 importance of $h_t$ w.r.t. $h_{t'}$
• Attention-focused representation $l_t$ of token $x_t$ given by:

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

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

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

\[ y^* = \arg\max_y \Pr(y|x; \Psi) \]
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

<table>
<thead>
<tr>
<th></th>
<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 O</td>
<td>B E</td>
<td>O O</td>
<td>B I E</td>
<td>E O</td>
<td></td>
<td></td>
<td></td>
<td>B O</td>
<td></td>
</tr>
<tr>
<td>B O</td>
<td>B O</td>
<td>O O</td>
<td>O O O</td>
<td>O O B</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tag flips = 4

• 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
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’