OpenTag: Open Attribute Value Extraction from Product Profiles

Guineng Zheng*, Subhabrata Mukherjee[Δ], Xin Luna Dong[Δ], FeiFei Li*
[Δ]Amazon.com, *University of Utah
{subhomj, lunadong}@amazon.com, {guineng, lifeifei}@cs.utah.edu

KDD 2018

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

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

- 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

Extracting Structured Values from Unstructured Text

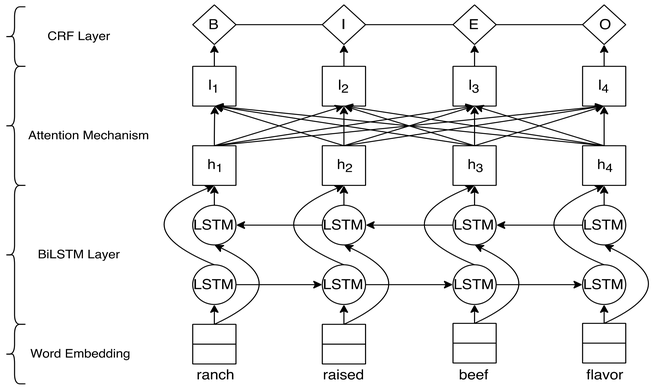
- Given a product title "Variety Pack Fillet Mignon and Porterhouse Steak Dog Food (12 Count)": OpenTag extracts attribute values "size = 12 count" and "flavor = {Porterhouse Streak, Fillet Mignon}"
  - It extracts *structured multi-word* values, and *multiple values* for an attribute
- Sequence Tagging: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

Sequence	duck	,	fillet	mignon	and	ranch	raised	lamb	flavor
BIOE	B	O	B	E	O	B	I	E	O
UBIOE	U	O	B	E	O	B	I	E	O
IOB	B	O	B	I	O	B	I	I	O

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: $h_t = \sigma([\overrightarrow{h_t}, \overleftarrow{h_t}])$
  - BiLSTM captures sequential nature of tokens but not tags
- Conditional Random Field (CRF) considers sequential nature of tags to extract coherent attribute values: $\Pr(y|x;\Psi) \propto \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \psi_k f_k(y_{t-1}, y_t, \langle l_t \rangle)\right)$
- Attention mechanism captures importance of tags for tokens, important concepts, and generates interpretable explanation for its verdict: $l_t = \sum_{t'=1}^{n} \alpha_{t,t'} \cdot h_{t'}$

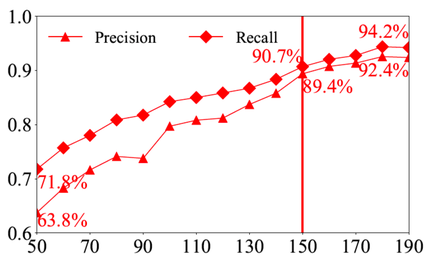
OpenTag: Architecture



OpenTag: Active Learning

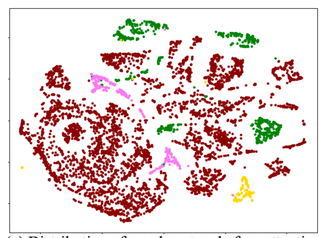
- Uses *active learning* to reduce manual annotation effort
  - To identify difficult-to-extract instances for the model, and ask for annotation
- Uses *tag flips as a* Query Strategy to identify difficult training instances
  - *Tag Flip:* Change in tag of a token across successive epochs during training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

$$Q^{tf}(x) = \sum_{e=1}^{E} \sum_{t=1}^{n} \mathcal{I}(y_t^*(\Psi^{(e-1)}) \neq y_t^*(\Psi^{(e)}))$$

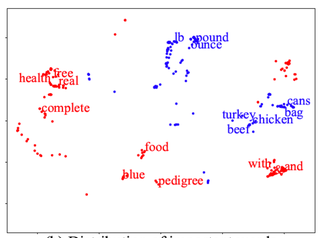


Datasets/Attribute	Models	Precision	Recall	Fscore
Dog Food: Title Attribute: Flavor	BiLSTM	83.5	85.4	84.5
	BiLSTM-CRF	83.8	85.0	84.4
	OpenTag	**86.6**	**85.9**	**86.3**
Camera: Title Attribute: Brand name	BiLSTM	94.7	88.8	91.8
	BiLSTM-CRF	91.9	**93.8**	92.9
	OpenTag	**94.9**	93.4	**94.1**
Detergent: Title Attribute: Scent	BiLSTM	81.3	82.2	81.7
	BiLSTM-CRF	**85.1**	82.6	83.8
	OpenTag	84.5	**88.2**	**86.4**
Dog Food: Description Attribute: Flavor	BiLSTM	57.3	58.6	58
	BiLSTM-CRF	62.4	51.5	56.9
	OpenTag	**64.2**	**60.2**	**62.2**
Dog Food: Bullet Attribute: Flavor	BiLSTM	93.2	94.2	93.7
	BiLSTM-CRF	94.3	94.6	94.5
	OpenTag	**95.7**	**95.7**	**95.7**
Dog Food: Title Multi Attribute: Brand, Flavor, Capacity	BiLSTM	71.2	67.4	69.3
	BiLSTM-CRF	72.9	67.3	70.1
	OpenTag	**76.0**	**68.1**	**72.1**

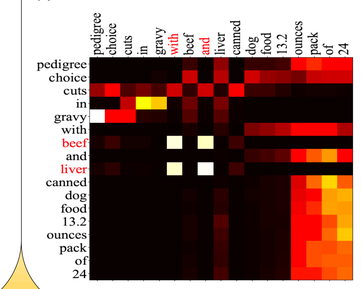
Interpretation via Attention



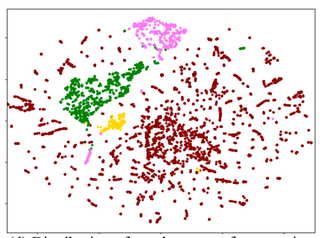
(a) Distribution of word vectors before attention



(b) Distribution of important words



(c) Heatmap of attention matrix



(d) Distribution of word vectors after attention