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

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

Brand

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

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

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

Open World Assumption

- No Predefined Attribute Value
- New Attribute Value Discovery

1. beef flavor
2. lamb flavor
3. venison flavor

Prior Work and Our Contributions

	Open World Assumption	No Lexicon, No Hand-crafted Features	Active Learning
Ghani et al. 2003, Putthividhya et al. 2011, Ling et al. 2012, Petrovski et al. 2017	✗	✗	✗
Huang et al. 2015, Kozareva et al. 2016	✓	✗	✗
Kozareva et al. 2016, Lample et al. 2016, Ma et al. 2016	✓	✓	✗
OpenTag (this work)	✓	✓	✓

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

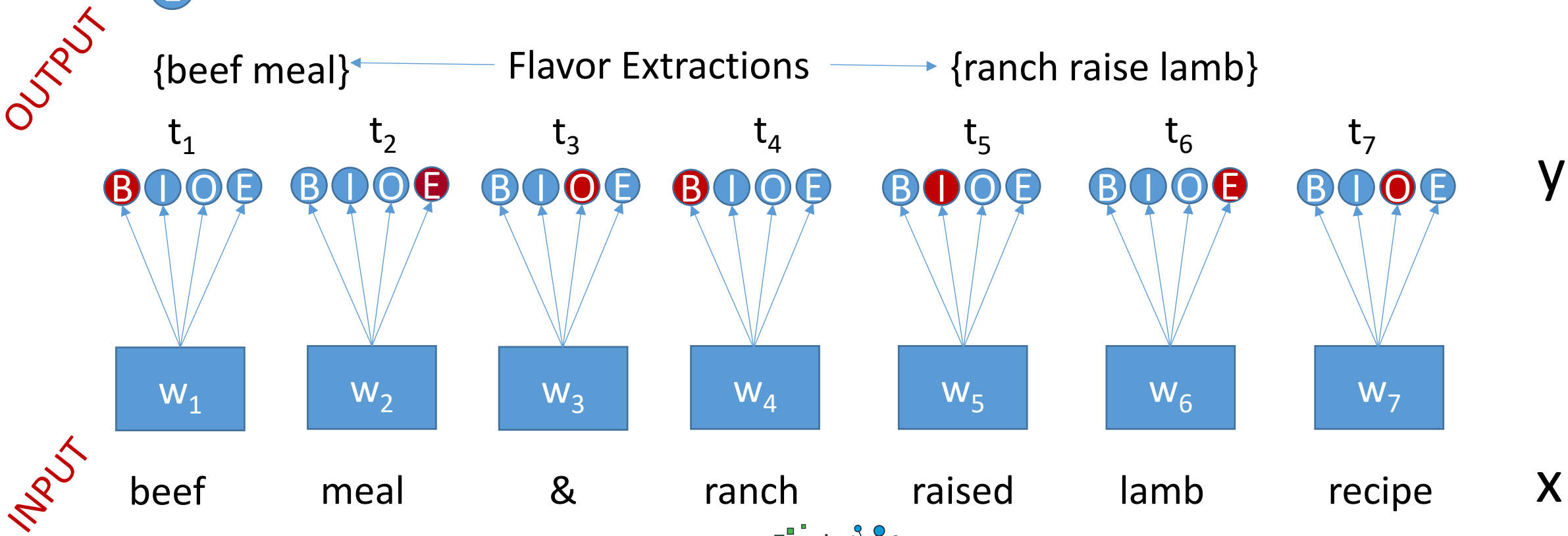
Input Product Profile			**Output** Extractions		
Title	Description	Bullets	Flavor	Brand	...
CESAR Canine Cuisine Variety Pack Filet Mignon & Porterhouse Steak Dog Food (Two 12-Count Cases)	A Delectable Meaty Meal for a Small Canine Looking for the right food … This delicious dog treat contains tender 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	

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

$x=\{w_1,w_2,\ldots,w_n\}$ input sequence

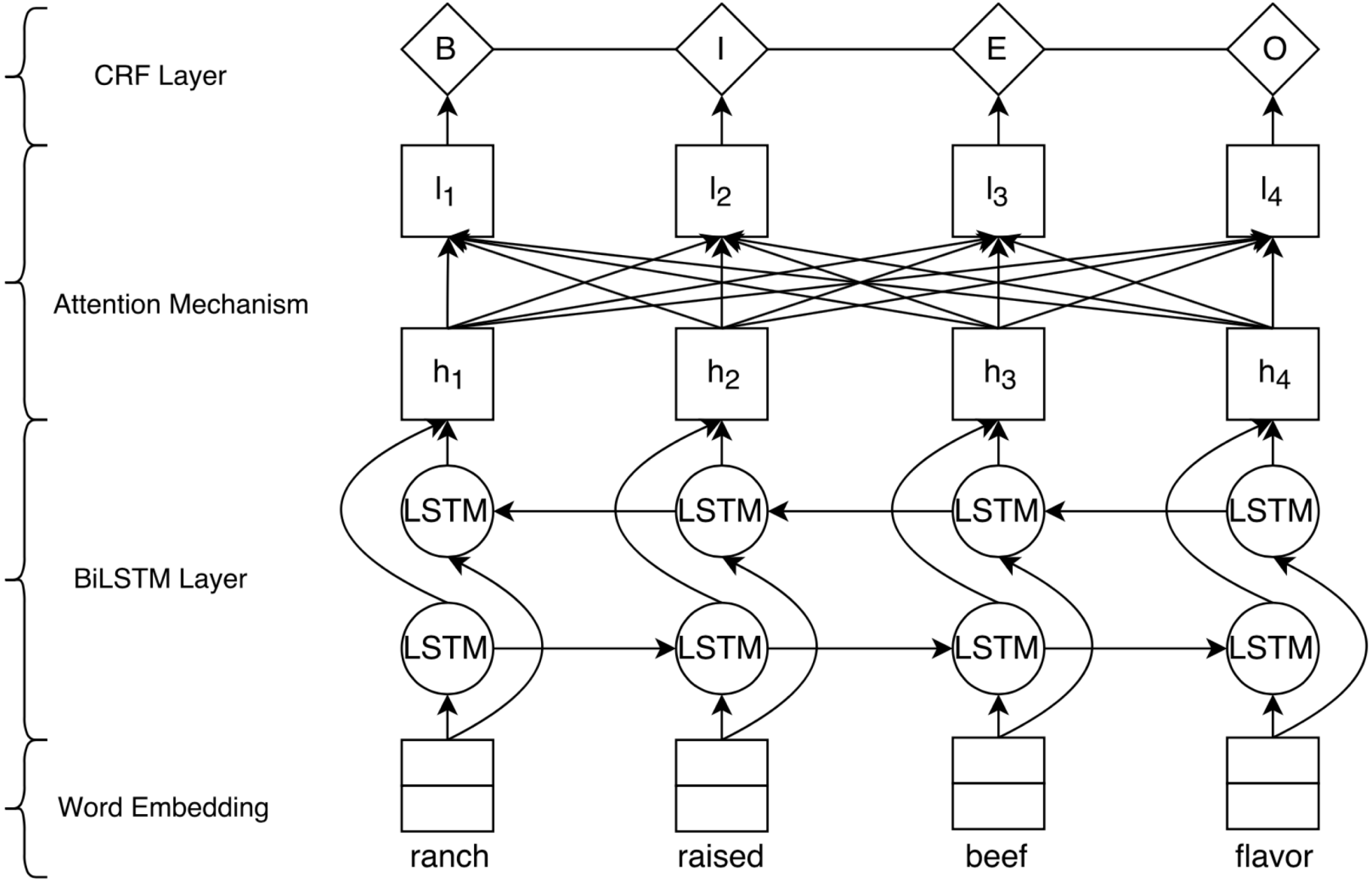
$y=\{t_1,t_2,\ldots,t_n\}$ tagging decision



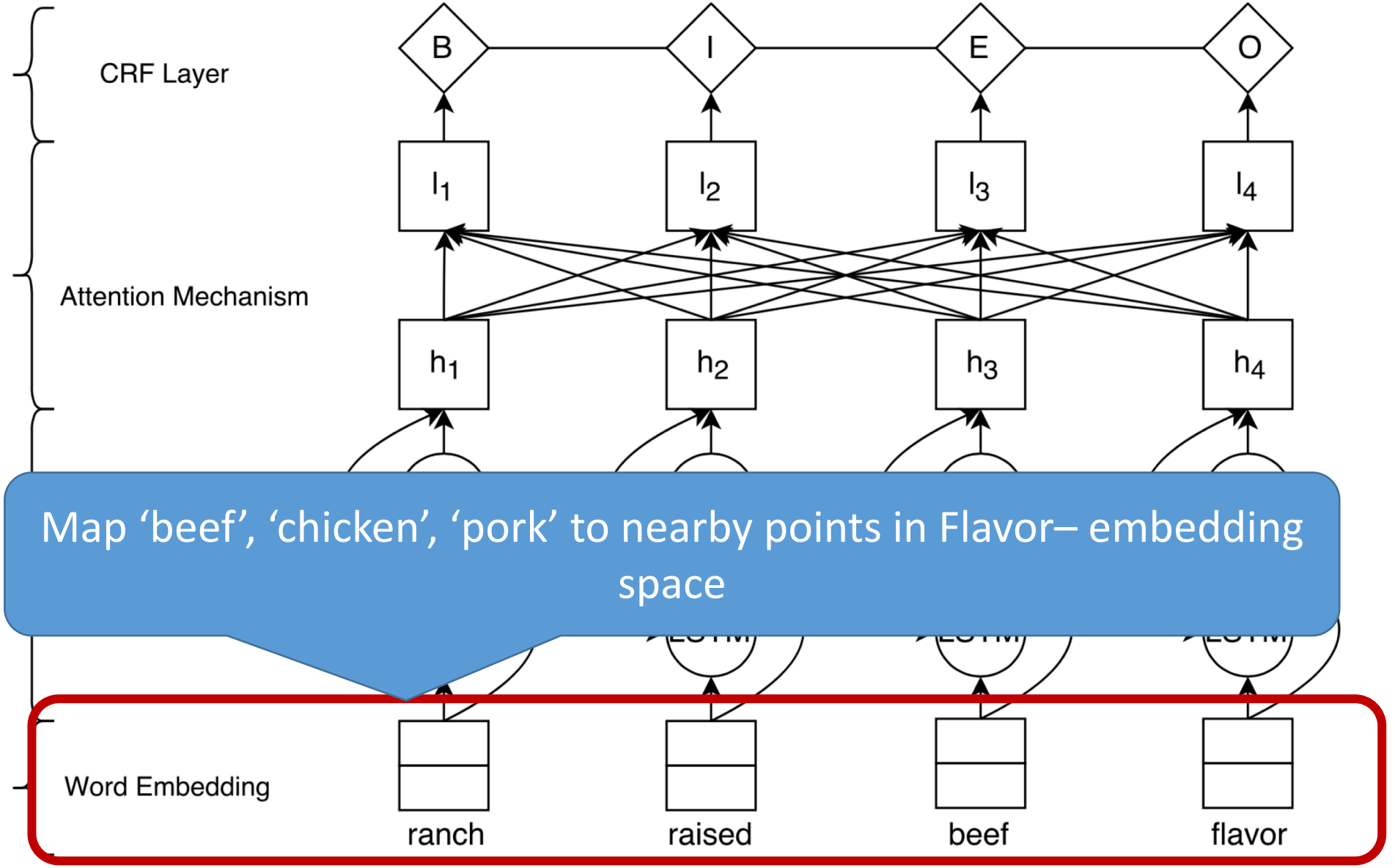
Outline

- Introduction
- Models
  - BiLSTM
  - BiLSTM + CRF
  - Attention Mechanism
  - OpenTag Architecture
- Active Learning

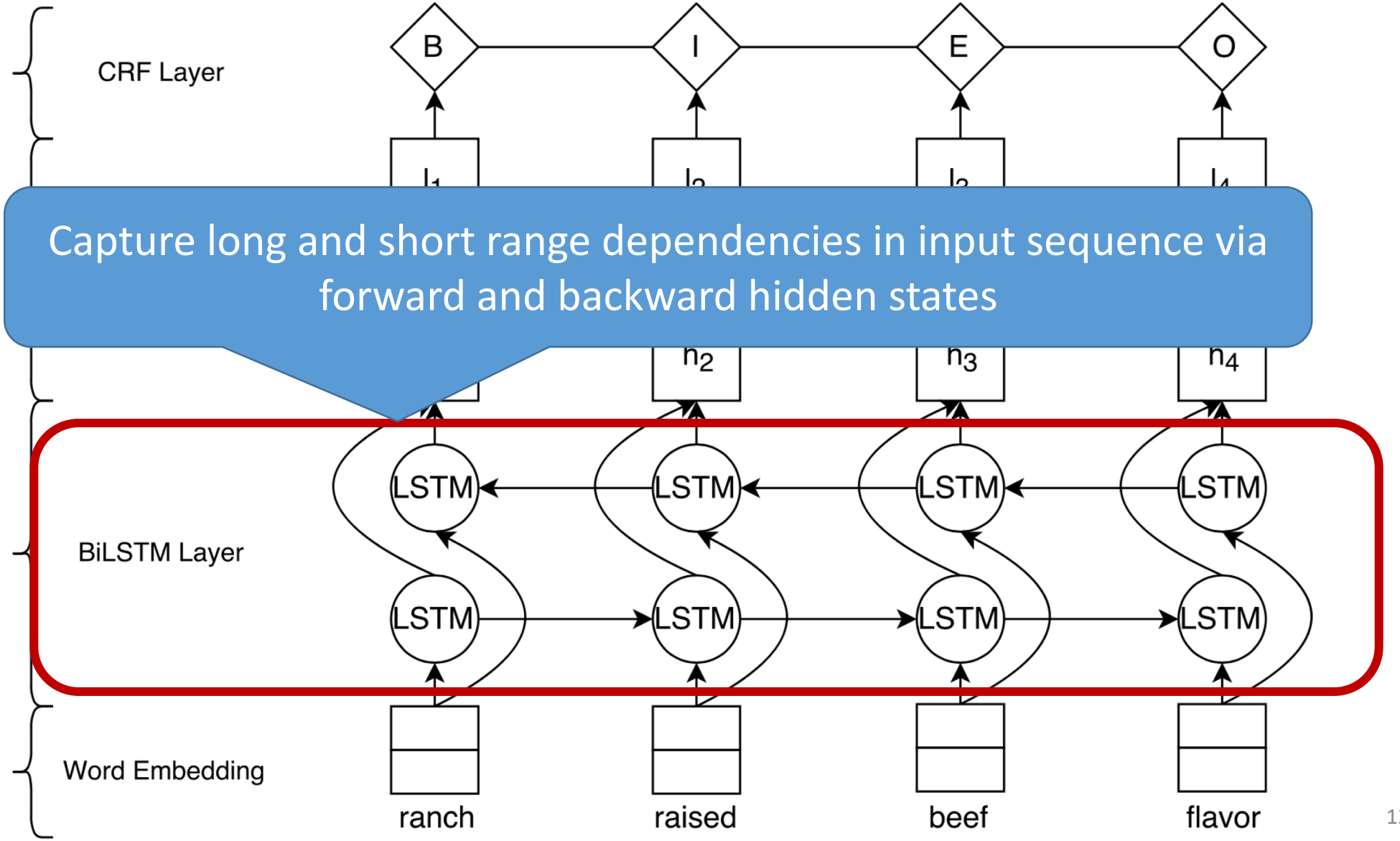
OpenTag Architecture



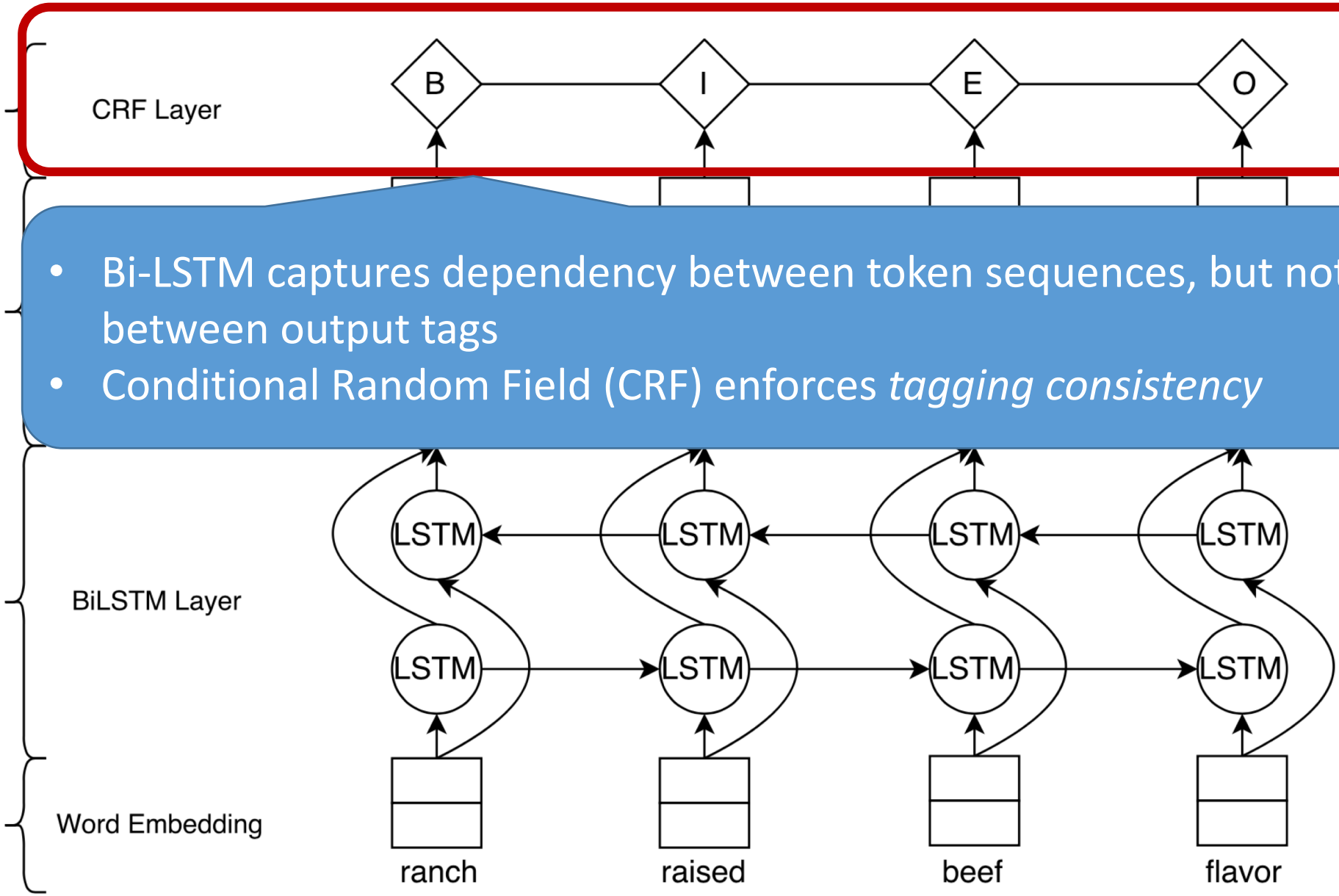
OpenTag Architecture (1/4): Word Embedding



OpenTag Architecture (2/4): Bidirectional LSTM

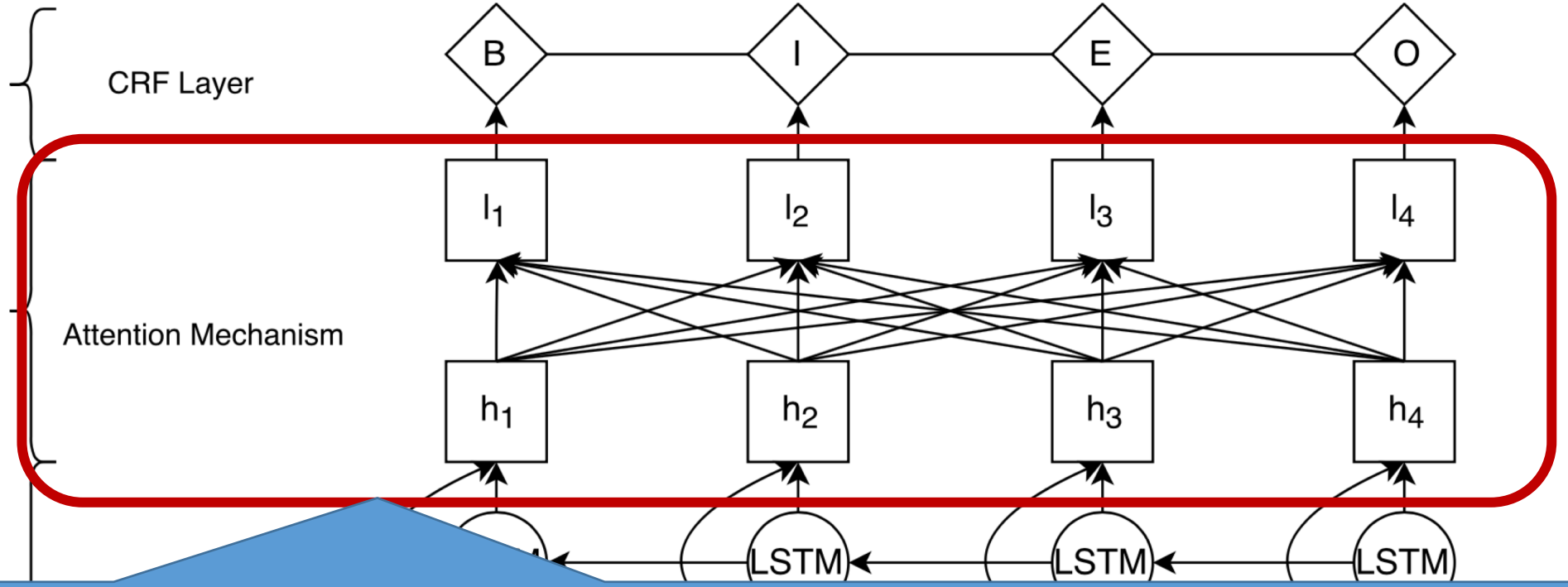


OpenTag Architecture (3/4): CRF



- Bi-LSTM captures dependency between token sequences, but not between output tags
- Conditional Random Field (CRF) enforces *tagging consistency*

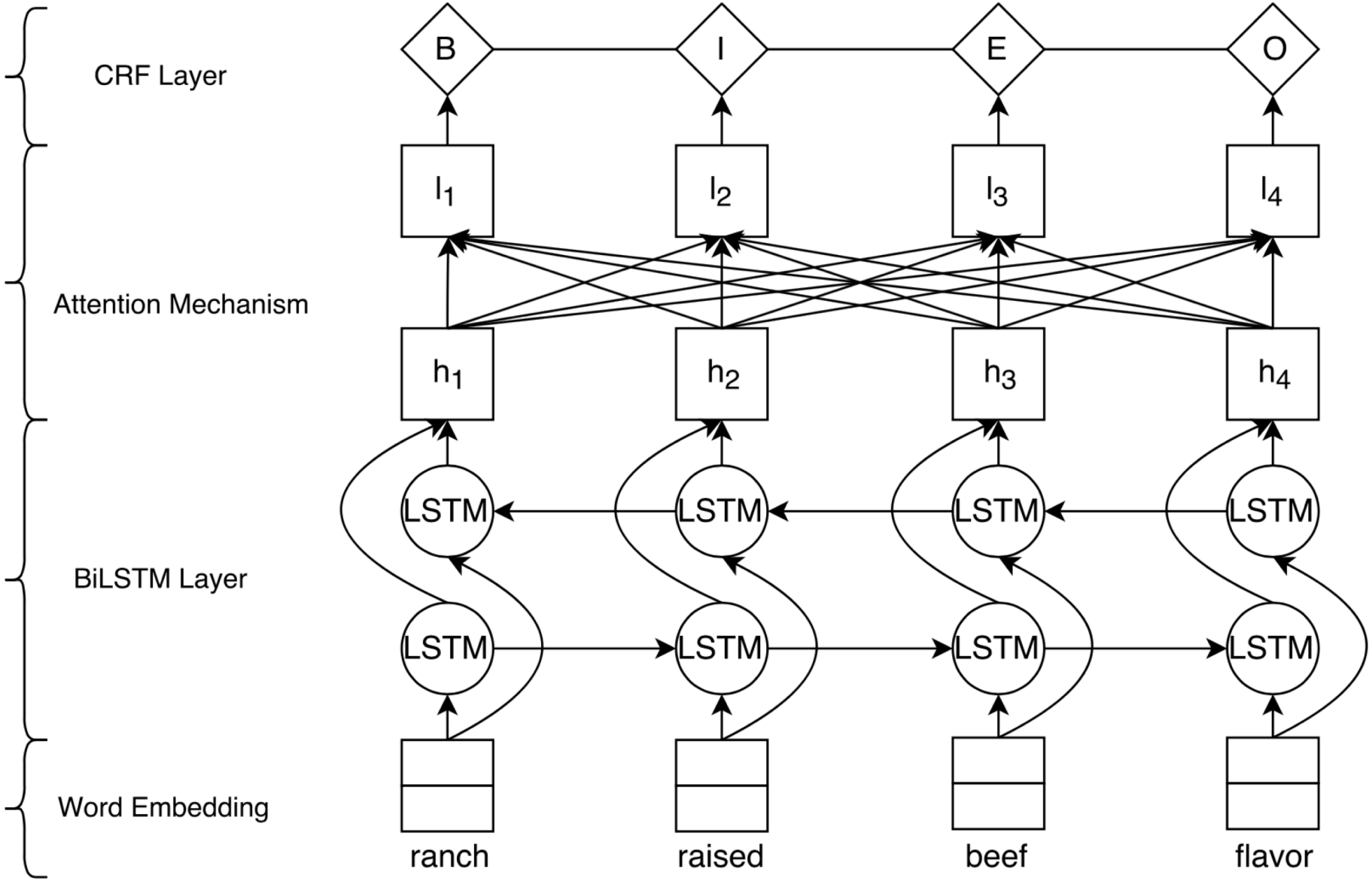
OpenTag Architecture (4/4): Attention



- 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



Experimental Discussions: Datasets

Domain	Profile	Attribute	Training		Testing	
			Samples	Extractions	Samples	Extractions
Dog Food (DS)	Title	Flavor	470	876	493	602
Dog Food	Title	Flavor	470	716	493	762
	Desc	Flavor	450	569	377	354
	Bullet	Flavor	800	1481	627	1179
	Title	Brand	470	480	497	607
	Title	Capacity	470	428	497	433
	Title	Multi	470	1775	497	1632
Camera	Title	Brand	210	210	211	211
Detergent	Title	Scent	500	487	500	484

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**
[illegible]	BiLSTM	81.3	82.2	81.7
	BiLSTM-CRF	**85.1**	82.6	83.8
	OpenTag	84.5	**88.2**	**86.4**
[illegible] 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**

Overall, OpenTag obtains high F-score of 82.8%

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

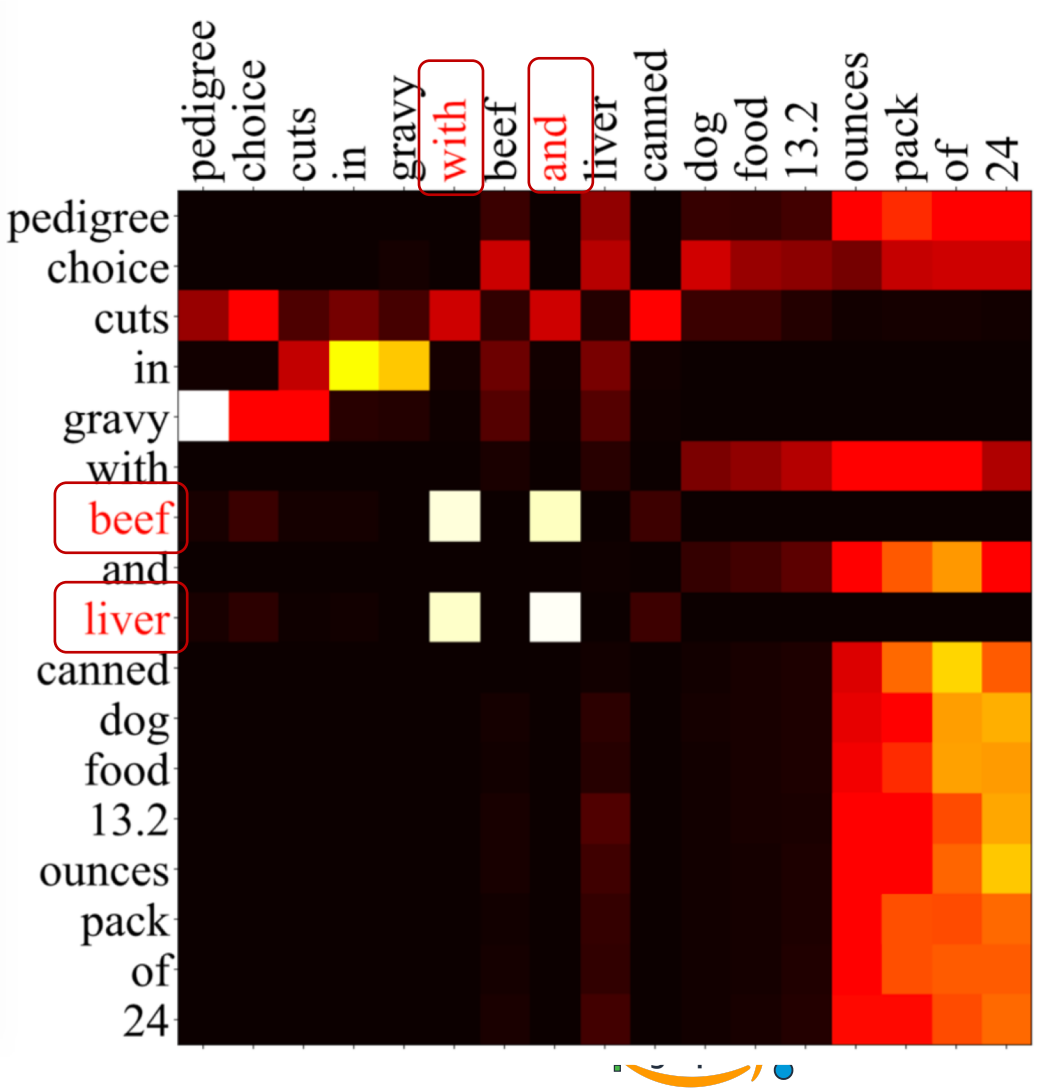
- Highest improvement in F-score of 5.3% over BiLSTM-CRF for product *descriptions*
  - However, less accurate than *titles*

OpenTag discovers new attribute-values not seen during training with 82.4% F-score

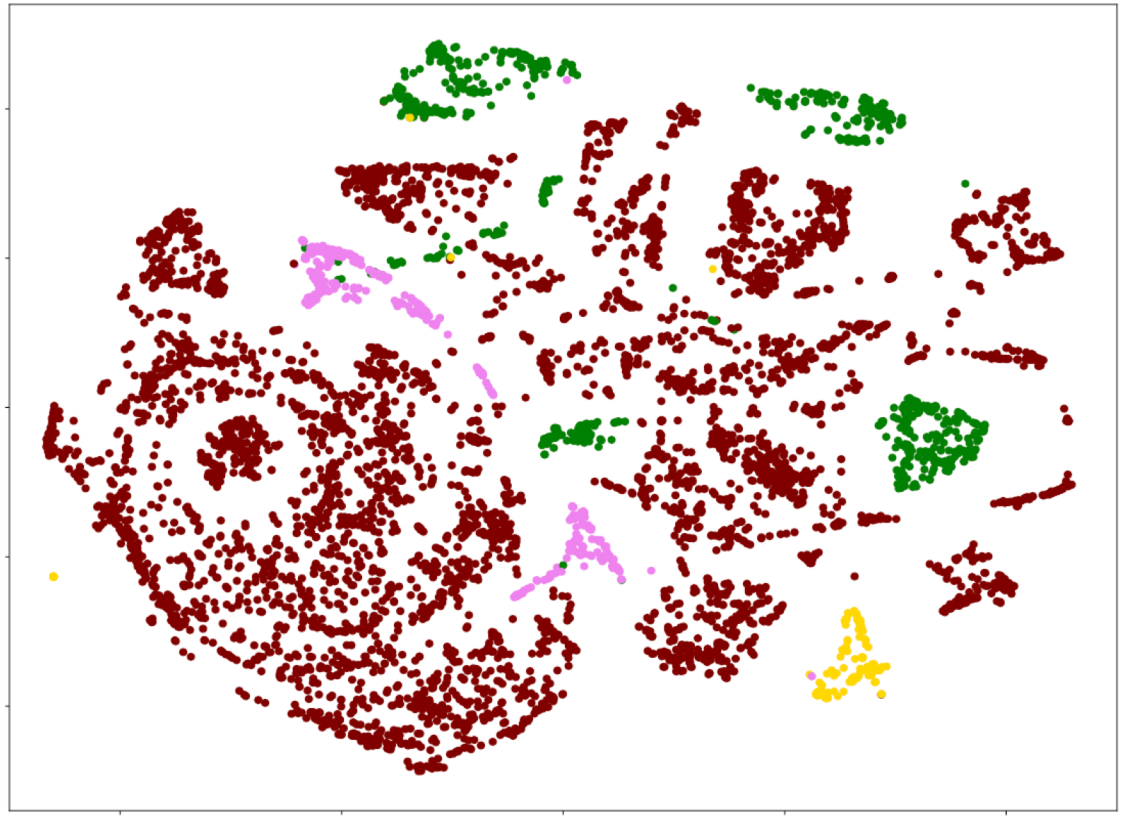
Train-Test Framework	Precision	Recall	F-score
Disjoint Split (DS)	83.6	81.2	82.4
Random Split	86.6	[illegible]	86.3

No overlap in attribute value between train and test splits

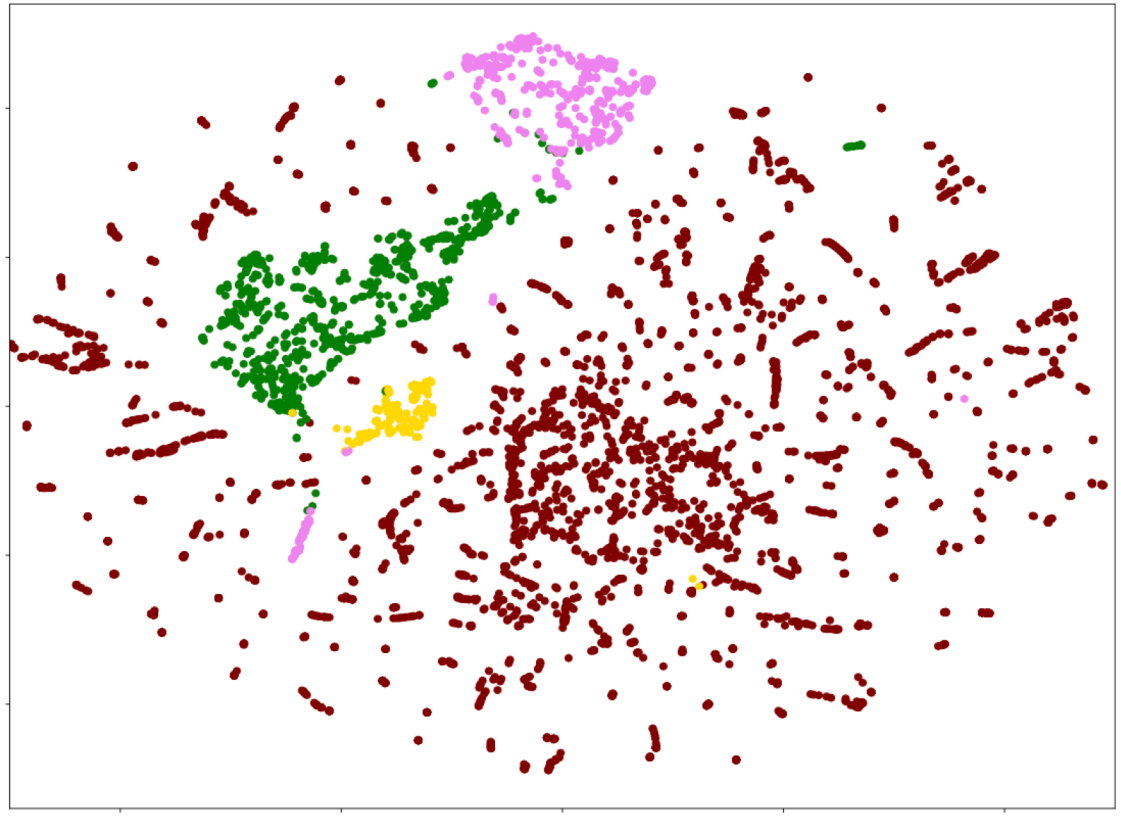
Interpretability via Attention



OpenTag achieves better concept clustering

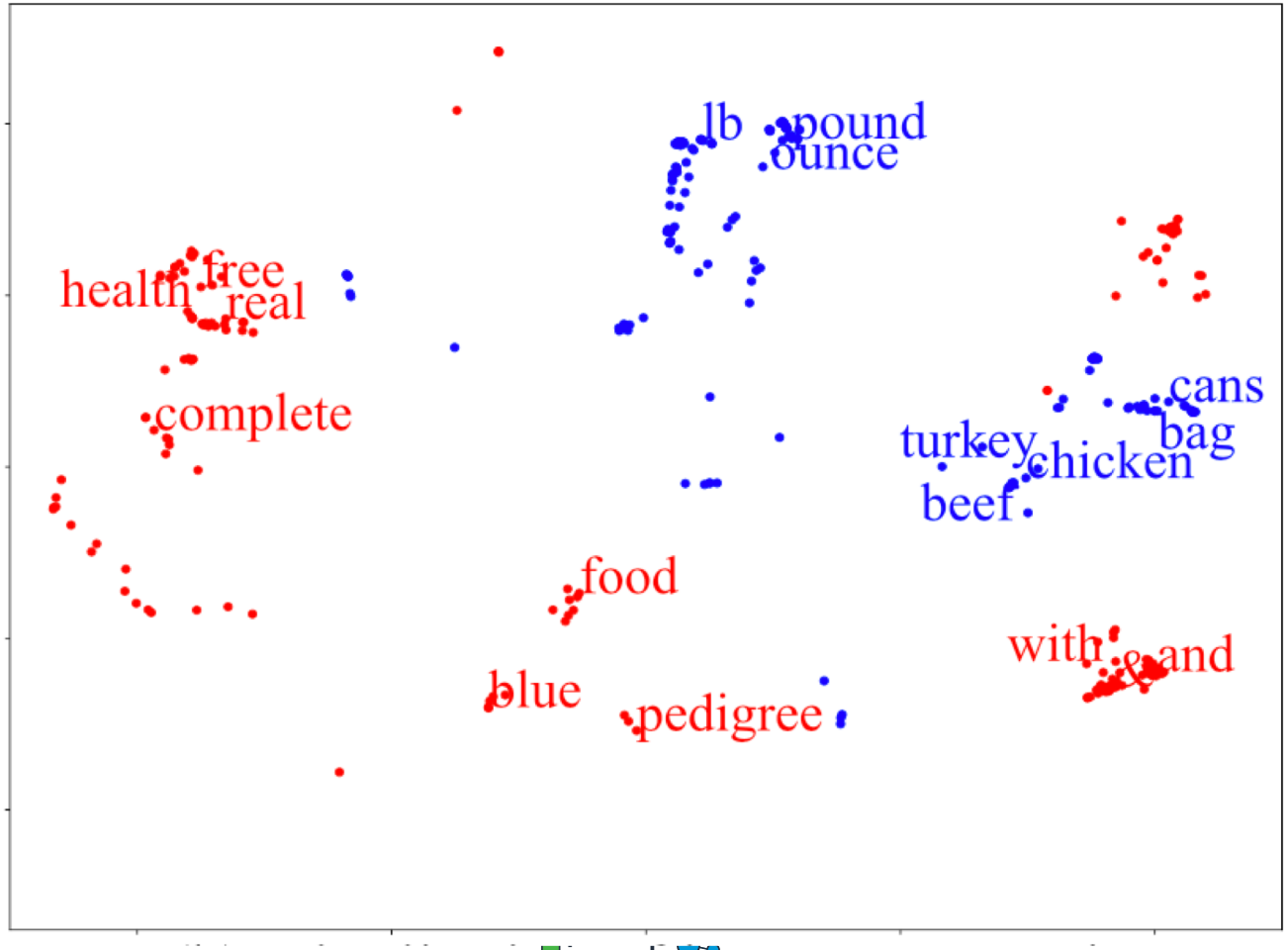


Distribution of word vectors before attention



Distribution of word vectors after attention

Semantically related words come closer in the embedding space



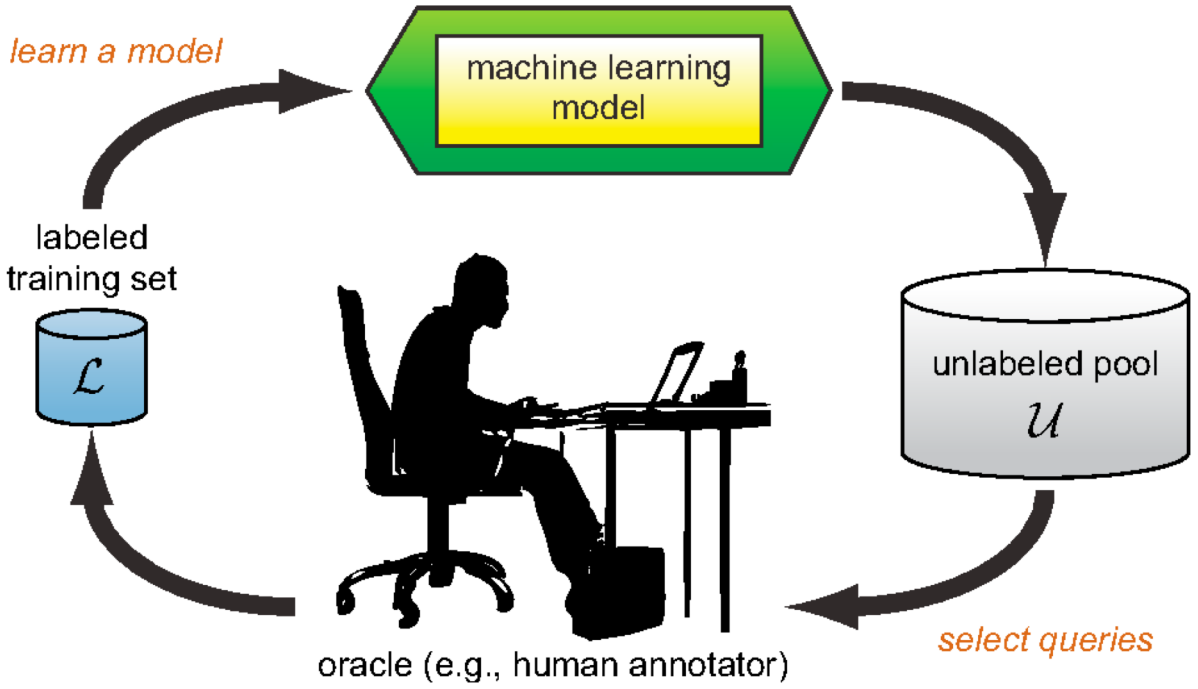
Outline

- 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

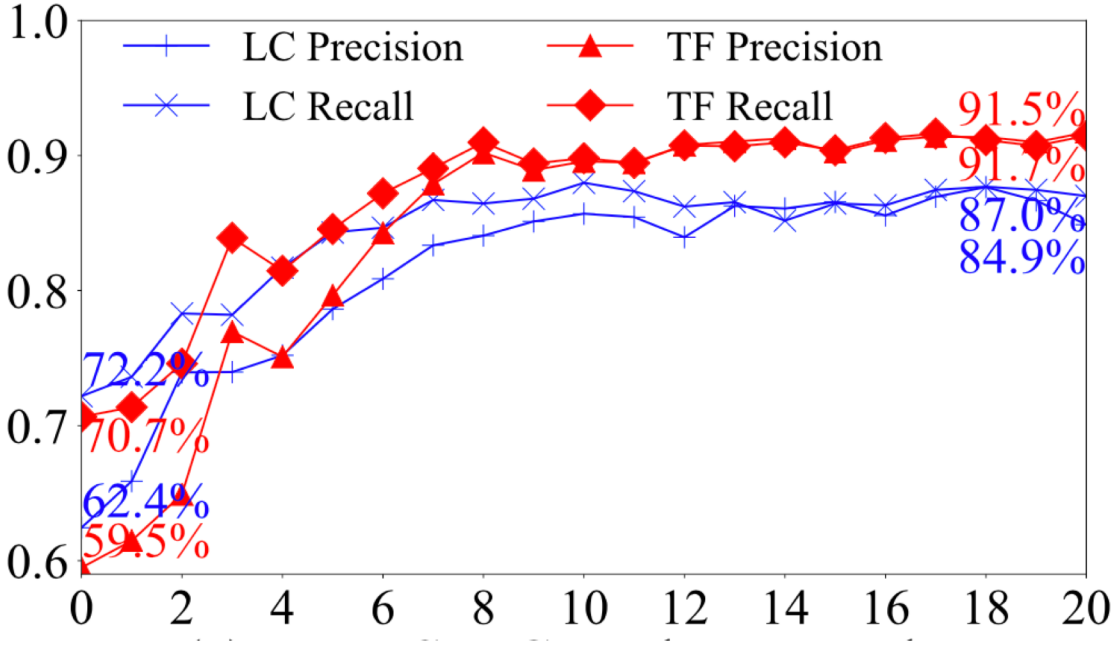
- Simulate a committee of OpenTag learners over multiple epochs
- Most informative sample => major disagreement among committee members for tags of its tokens across epochs
- Use *dropout mechanism* for simulating committee of learners

duck	,	fillet	mignon	and	ranch	raised	lamb	flavor
B	O	B	E	O	B	I	E	O
B	O	B	O	O	O	O	B	O

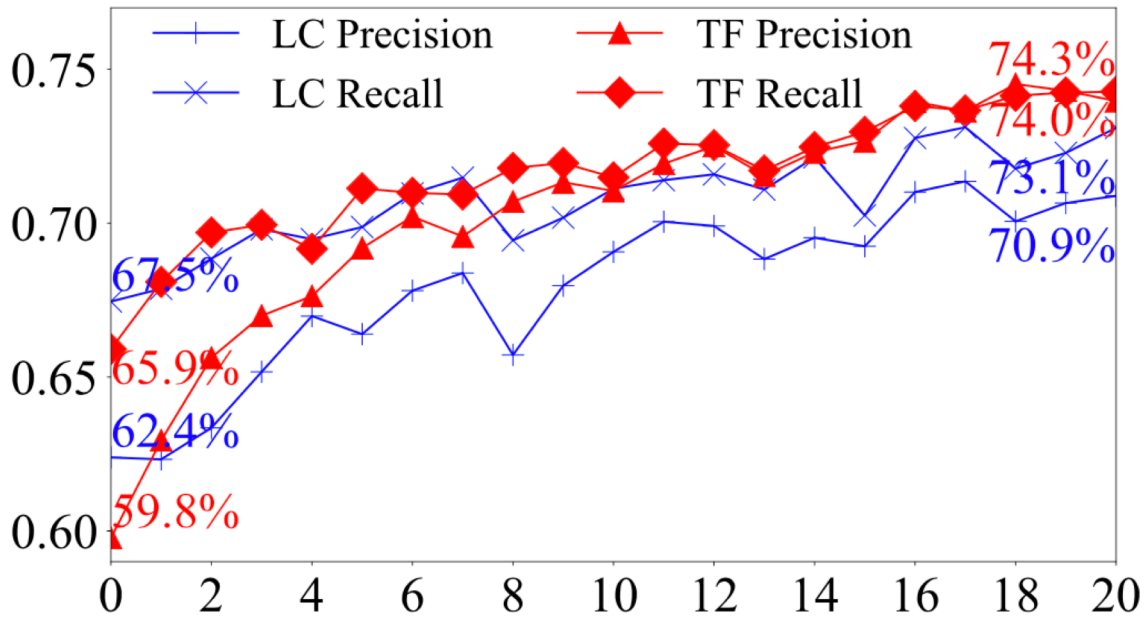
Tag flips = 4

- Most informative sample has *highest tag flips* across all the epochs

Tag Flip (red) better than Uncertainty Sampling (blue)

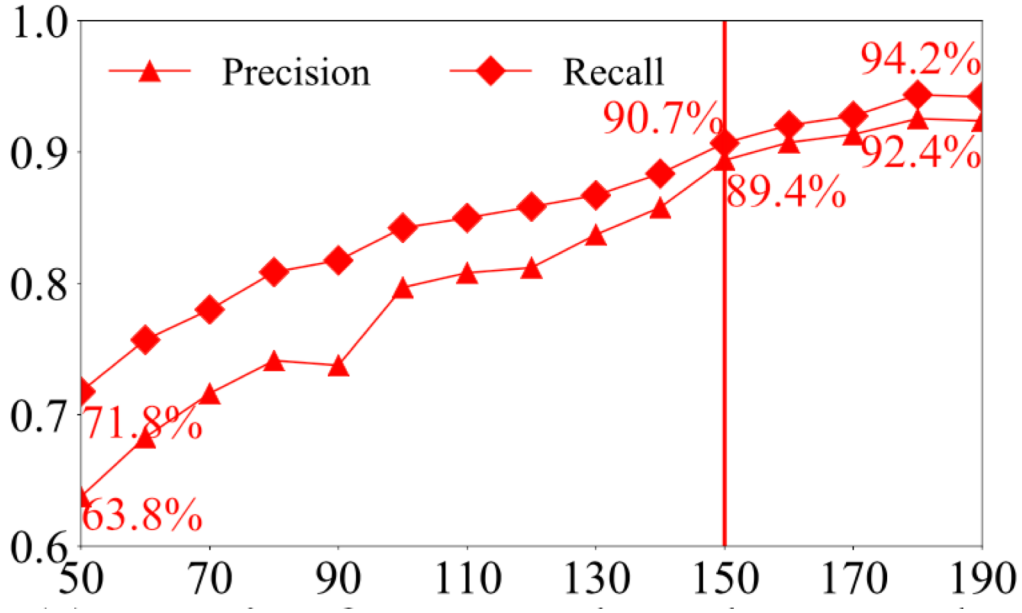


TF v.v. LC on detergent data

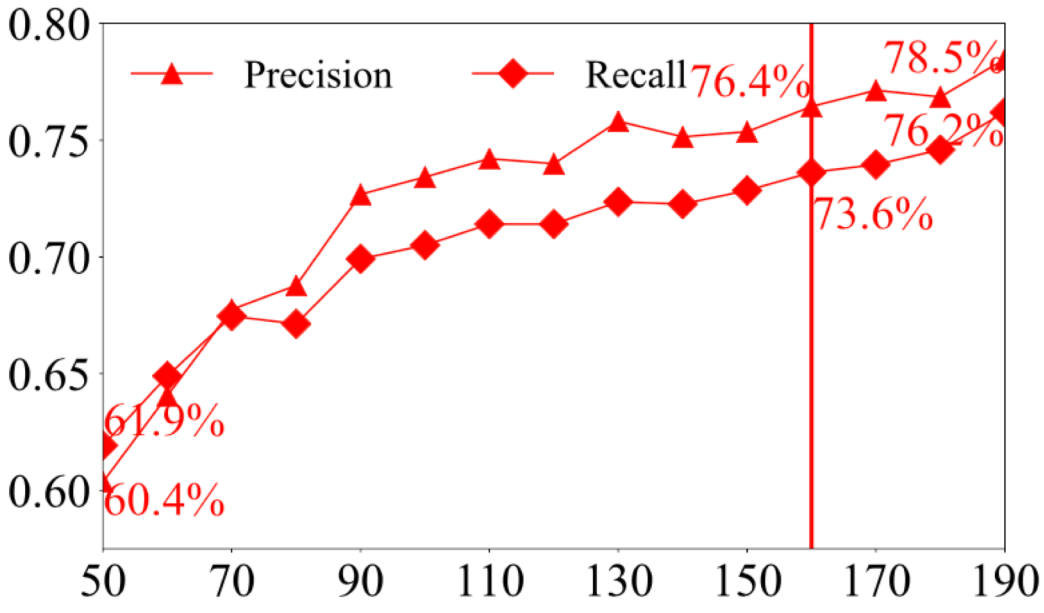


TF v.v. LC on multi extraction

OpenTag reduces burden of human annotation by 3.3x



Learning from scratch on detergent data



Learning from scratch on multi extraction

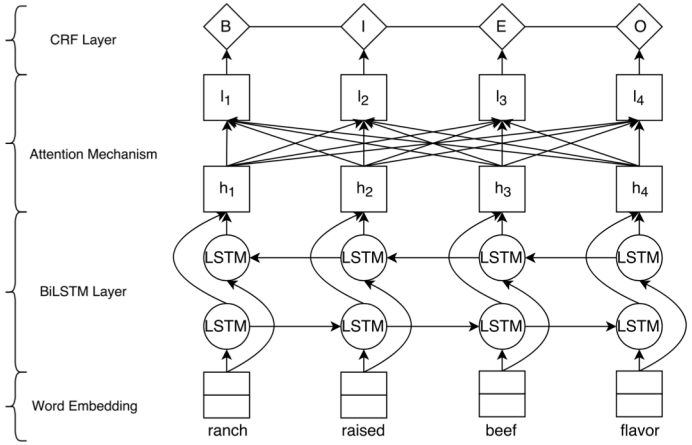
- OpenTag requires only 500 training samples to obtain > 90% P-R
- Active learning brings it down to 150 training samples to match similar performance

Production Impact

	Previous Coverage of Existing Production System (%)	OpenTag Coverage (%)	***Increase*** in Coverage (%)
Attribute_1	23	78	53
Attribute_2	21	72	45
Attribute_3	< 1	56	50
Attribute_4	< 1	49	48

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

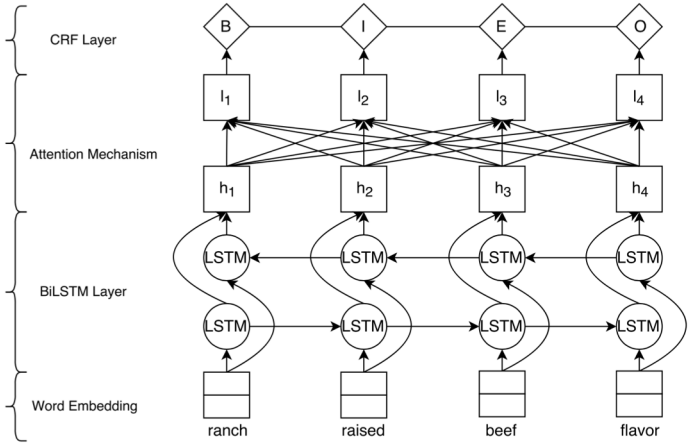
amazon

Thank you for your attention!

THE UNIVERSITY OF UTAH

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.



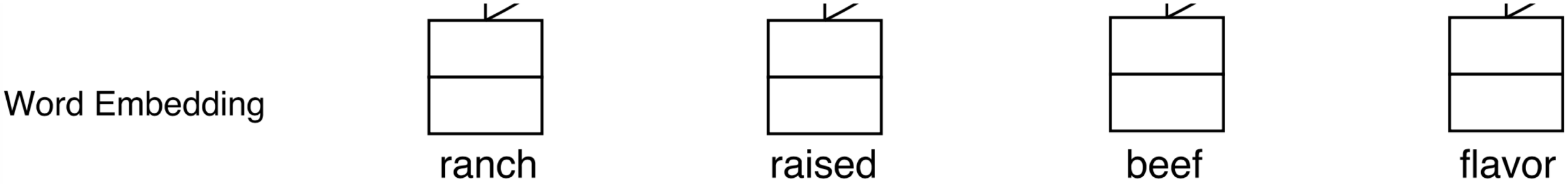
KDD 2018

product graph

Backup Slides

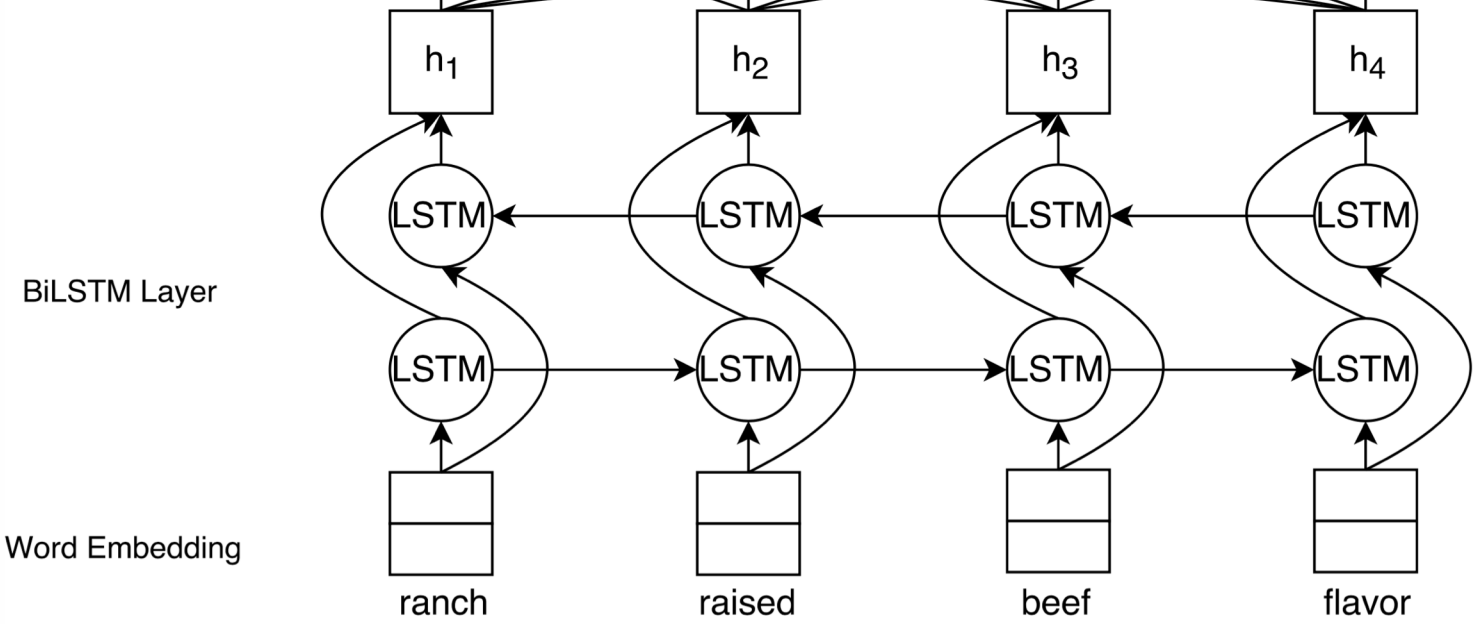
Word Embedding

- 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



amazon

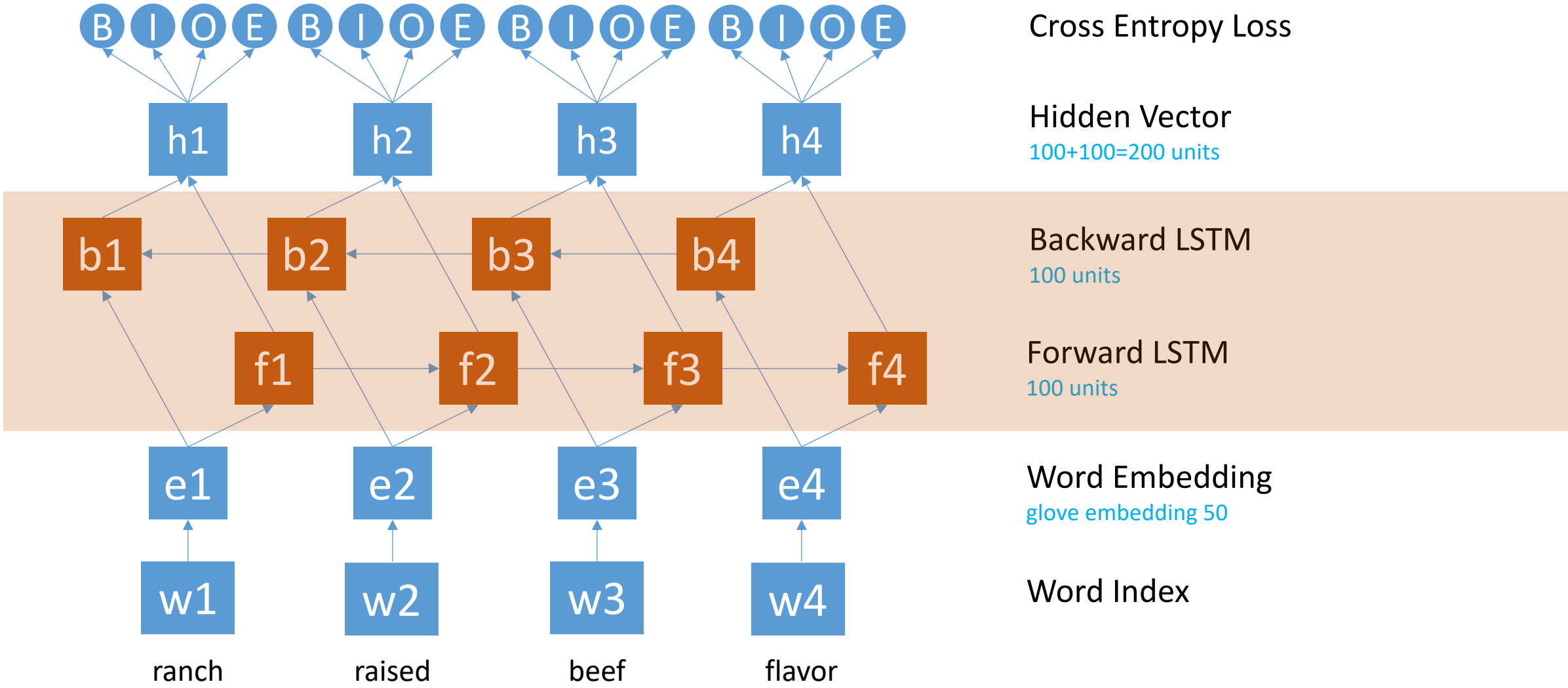
Bi-directional LSTM



- 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: $h_t = \boldsymbol{\sigma}([b_t, f_t])$

Bi-directional LSTM

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



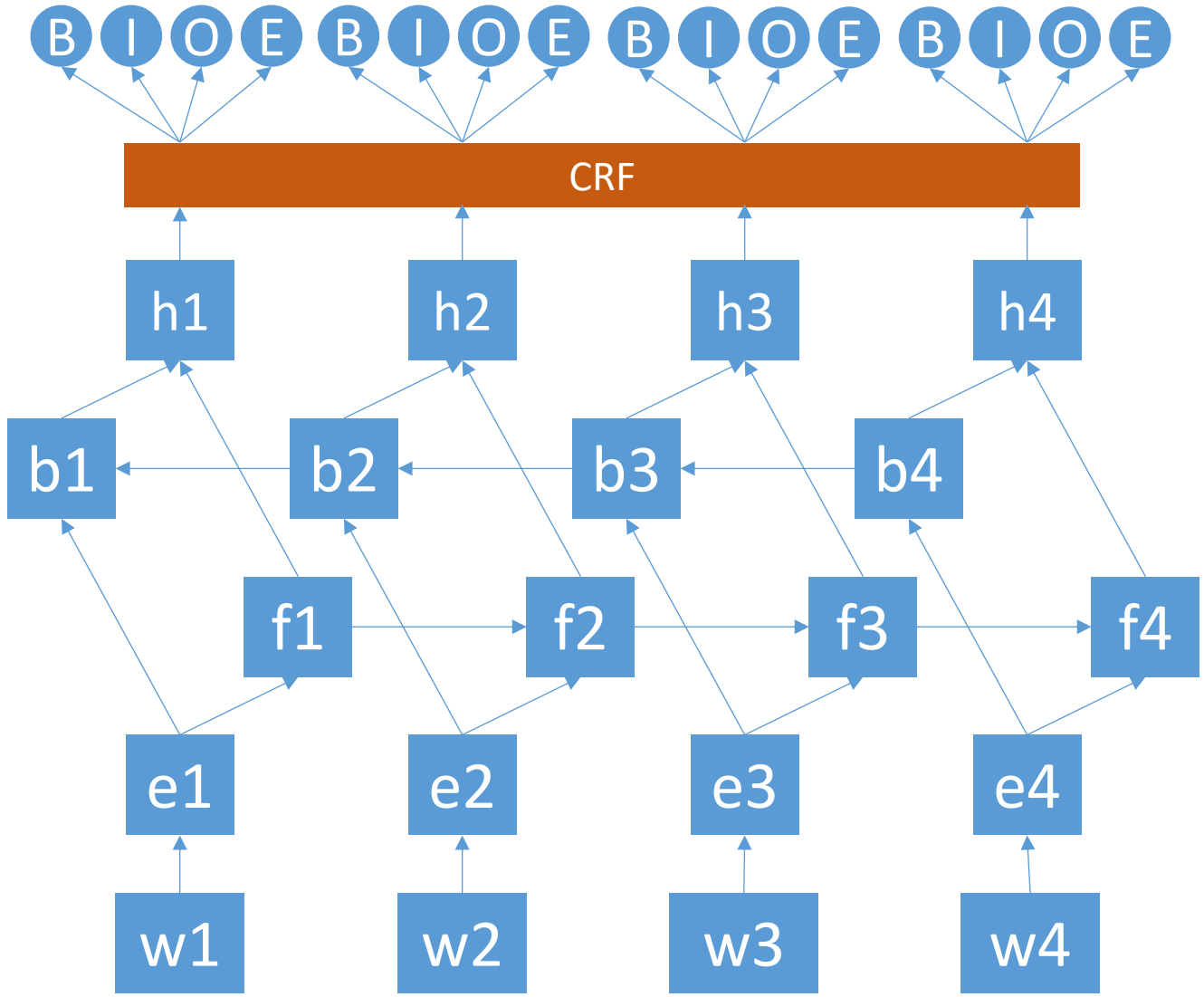
Conditional Random Fields (CRF)

- 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** = $\{x_1, x_2, \ldots, x_n\}$ with tags **y** = $\{y_1, y_2, \ldots, y_n\}$: 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

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



Cross Entropy Loss

Conditional Random Fiel

CRF feature space formed by Bi-LSTM hidden states

Forward LSTM

100 units

Embedding

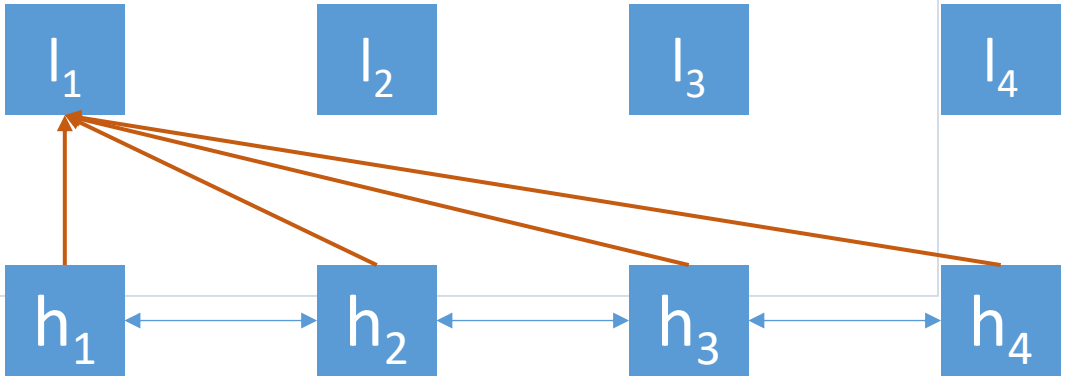
glove embedding 50

Word Index

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)
  - $\alpha_{t,t'} \in \mathbf{A}$ captures importance of h_t w.r.t. $h_{t'}$
- Attention-focused representation l_t of token x_t given by:

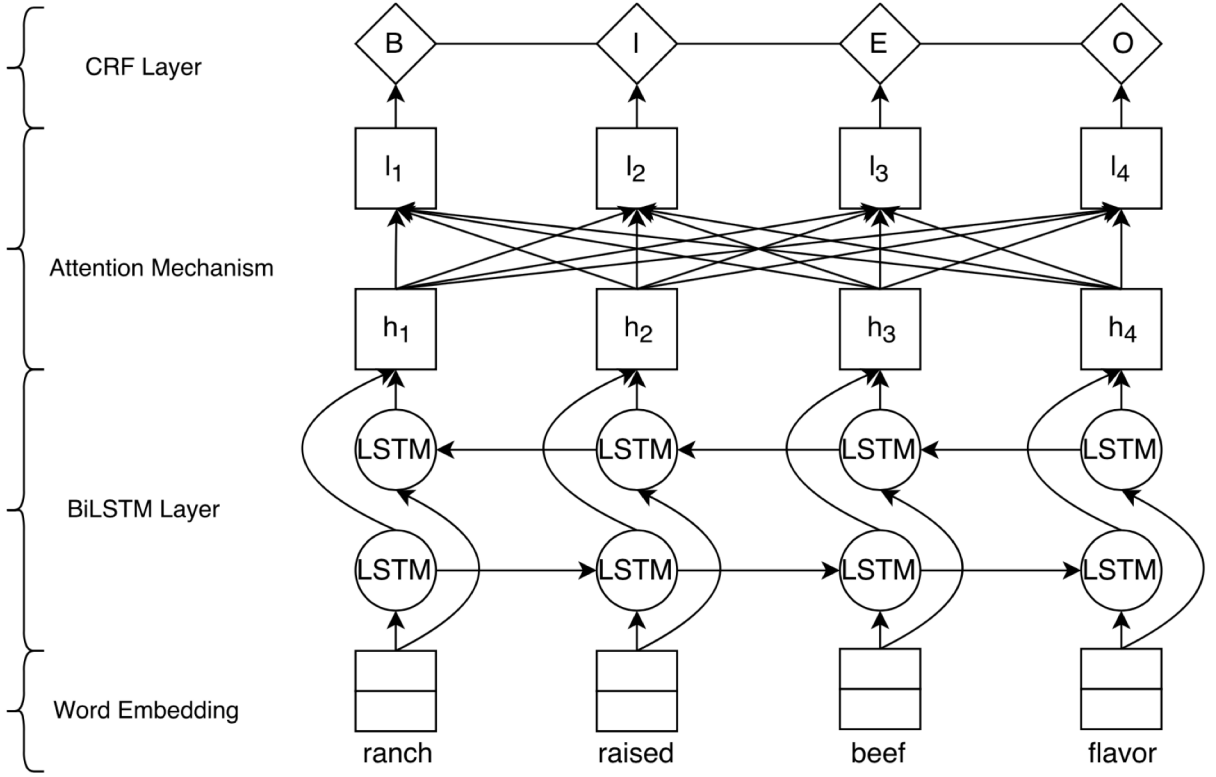
$$l_t = \sum_{t'=1}^{n} \alpha_{t,t'} \cdot h_{t'}$$



Final Classification

CRF feature space formed by attention-focused representation of hidden states

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



Maximize log-likelihood of joint distribution

$$L(\Psi) = \sum_{i=1}^{m} \log \Pr(\mathbf{y}_i|\mathbf{x}_i;\Psi)$$

Best possible tag sequence with highest conditional probability

$$y^* = \text{argmax}_y \Pr(y|x;\Psi)$$

Uncertainty Sampling: Probability as Query Strategy

- Select instance with maximum uncertainty
  - Best possible tag sequence from CRF:

$$y^* = \text{argmax}_y \Pr(y|x;\Psi)$$

  - 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

duck	,	fillet	mignon	and	ranch	raised	lamb	flavor
B	O	B	E	O	B	I	E	O
B	O	B	O	O	O	O	B	O

Tag flips = 4

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

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

- Top *B* samples with the highest number of flips are manually annotated with tags

Multiple attribute values

- Predicting multiple attribute values **jointly**

Attribute	Precision	Recall	F-Score
Brand: Single	52.6	42.6	47.1
Brand: Multi	**58.4**	**44.7**	**50.6**
Flavor: Single	83.6	**81.2**	**82.4**
Flavor: Multi	**83.7**	77.5	80.5
Capacity: Single	81.5	86.4	83.9
Capacity: Multi	**87.0**	**87.2**	**87.1**

- Modify tagging strategy to have separate tag-set $\{B_a, I_a, O_a, E_a\}$ for each attribute 'a'