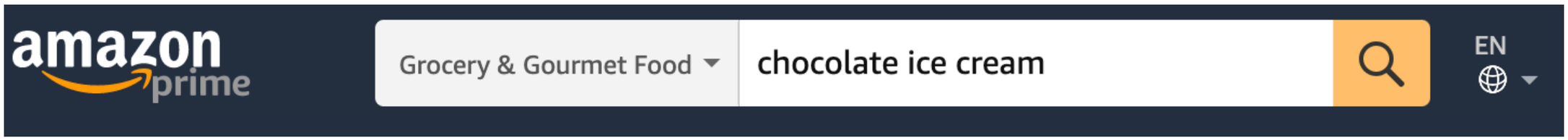
TXtract:
Taxonomy-Aware Knowledge Extraction for Thousands of Product Categories

Giannis Karamanolakis
Columbia University
gkaraman@cs.columbia.edu

Jun Ma, Xin Luna Dong
Amazon.com
{junmaa, lunadong}@amazon.com

Product Understanding
for Search and Question Answering

Product Understanding
for Search and Question Answering



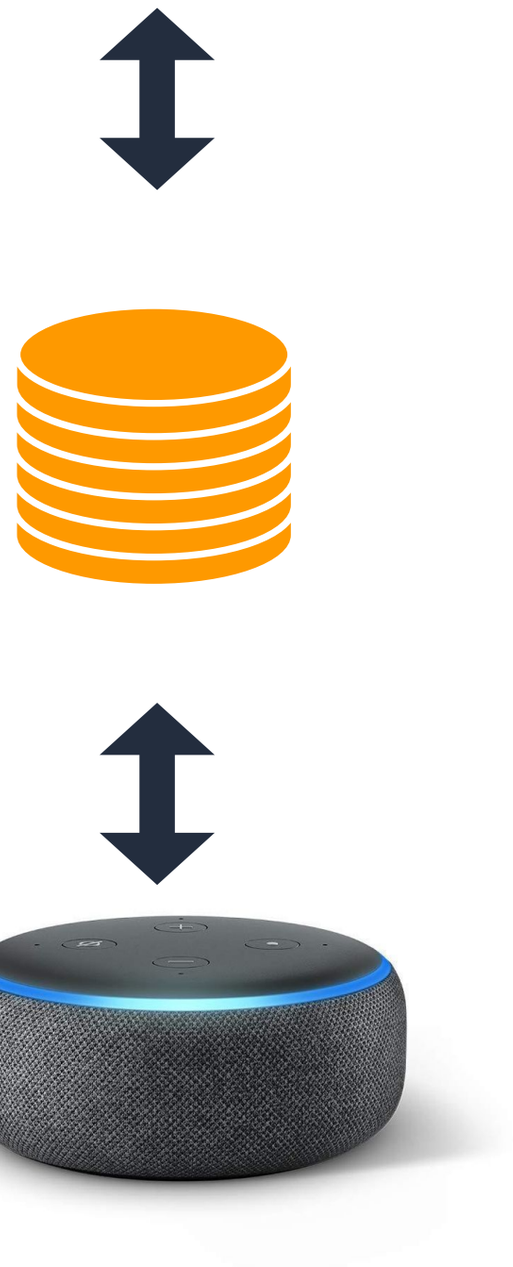
“Alexa, which shampoos contain argan oil?”

Need to Store Structured Knowledge About Products

flavor: “chocolate”

ingredients: “biotin”, “argan oil”, …

“Alexa, which shampoos contain argan oil?”



Understanding Values for Product Attributes

flavor: ???

Product catalog:

Product Attributes

Products

Product ID	Brand	Flavor	Size	Ingredients
B00FZHEGGW	Fage	Plain	35.3 oz	…
B0725VRRLP	Ben & Jerry’s	…	…	…
…	…	…	…	…

Understanding Values for Product Attributes

flavor: ???

Product catalog:

Product Attributes

Products

Product ID	Brand	Flavor	Size	Ingredients
B00FZHEGGW	Fage	Plain	35.3 oz	…
B0725VRRLP	Ben & Jerry's	???	???	…
…	…	…	…	…

(-) Issue: catalog is missing attribute values for many products

Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions

Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions

- **Previous work:** deep neural networks for **sequence tagging**

[Zheng et al., KDD'18]
[Xu et al., ACL'19]
[Rezk et al., ICDE'19]

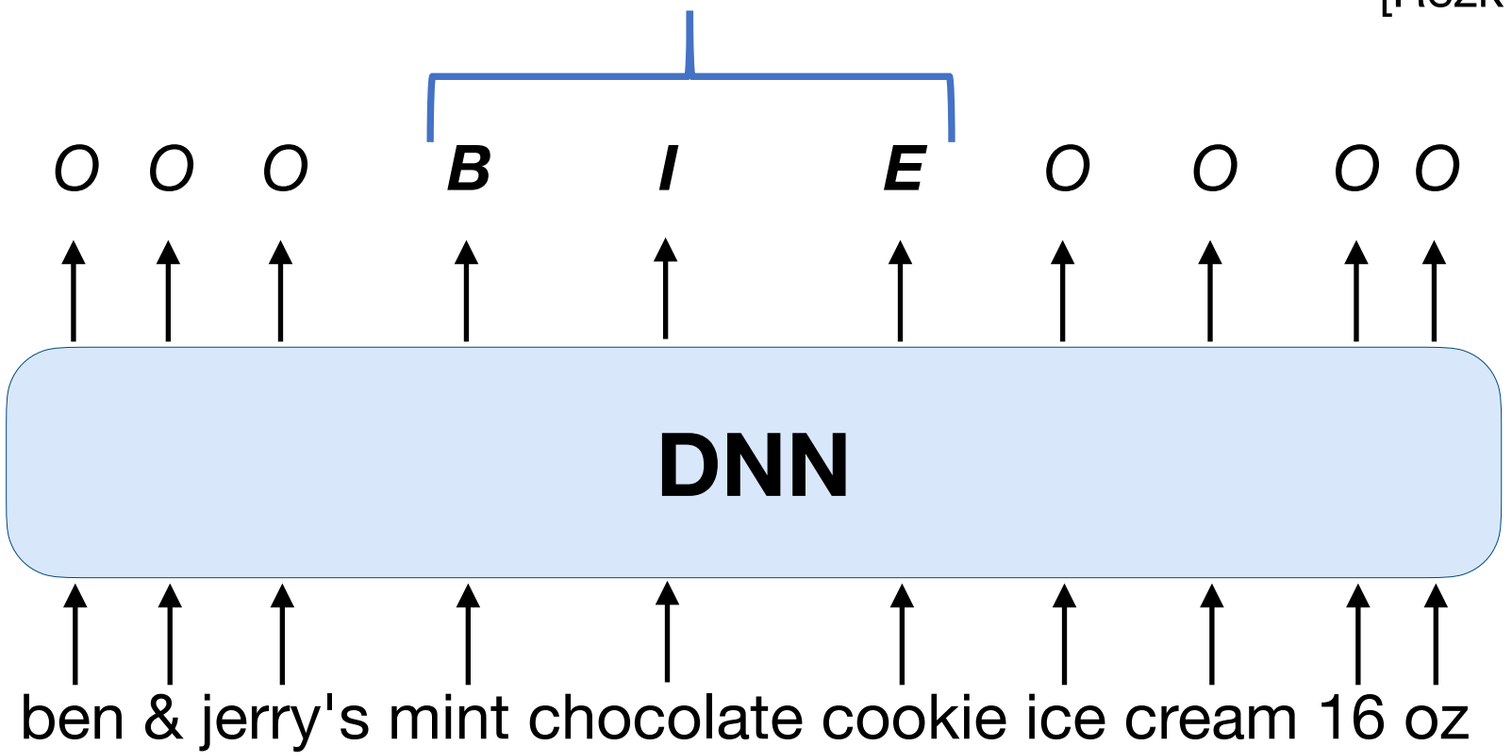
Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions
- **Previous work:** deep neural networks for **sequence tagging**

BIOE Tagging Example

extracted *flavor* value: “mint chocolate cookie”

[Zheng et al., KDD’18]
[Xu et al., ACL’19]
[Rezk et al., ICDE’19]



Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions
- **Previous work:** deep neural networks for **sequence tagging**
- **Limitations of previous work:**

(-) designed for a single category

[Zheng et al., KDD'18]

"Dog Food"
Atribute Values

DNN

"Dog Food"
Product Titles

[Xu et al., ACL'19]

"Sports"
Atribute Values

DNN

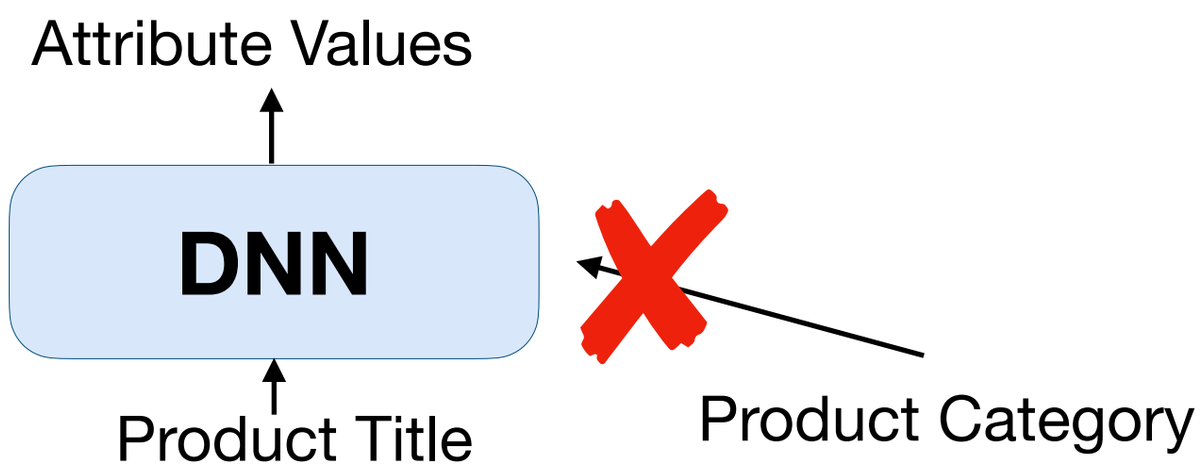
"Sports"
Product Titles

Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions
- **Previous work:** deep neural networks for **sequence tagging**
- **Limitations of previous work:**

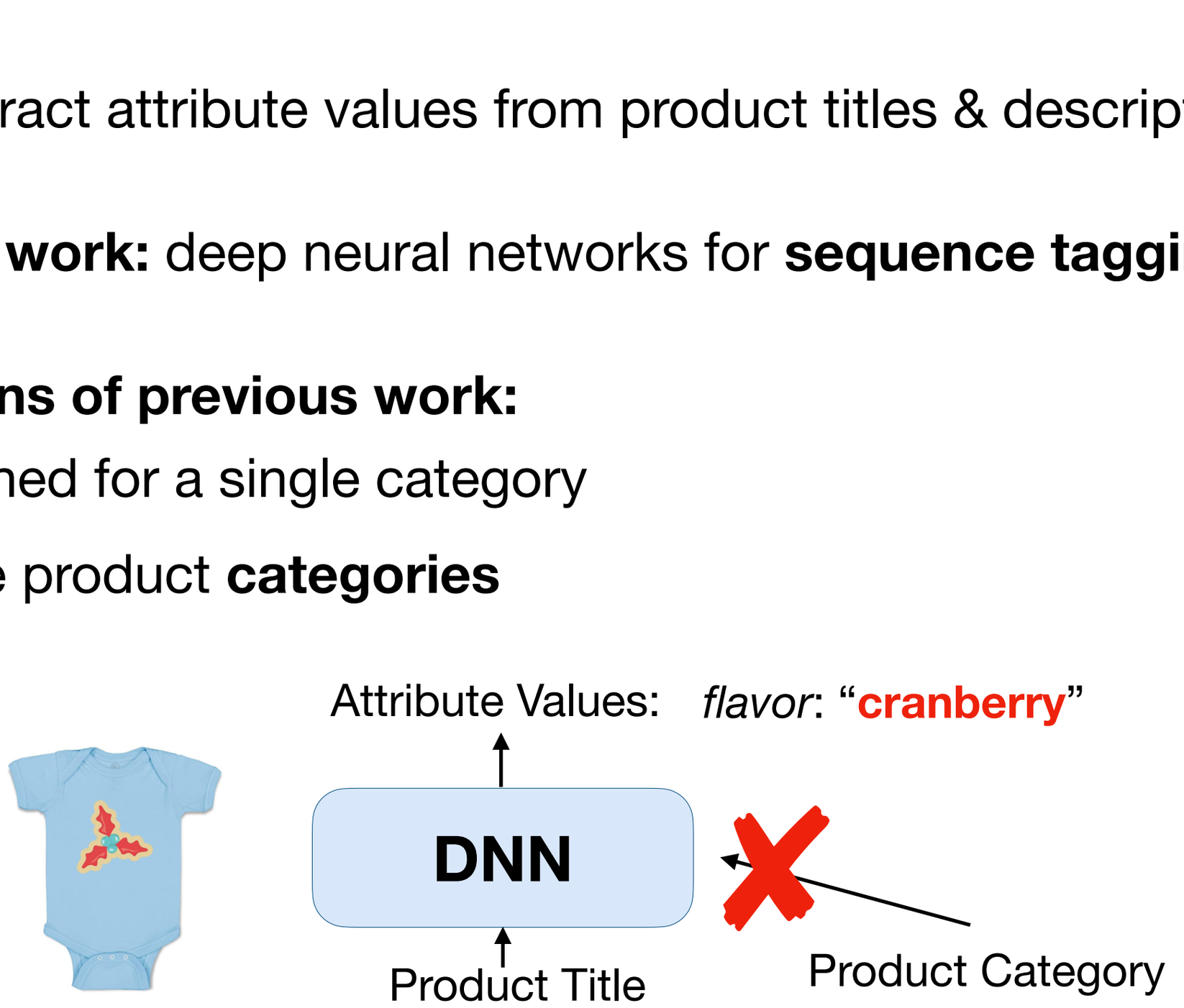
(-) designed for a single category

(-) ignore product **categories**



Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions
- **Previous work:** deep neural networks for **sequence tagging**
- **Limitations of previous work:**
  - **(-)** designed for a single category
  - **(-)** ignore product **categories**



*"Custom Boy & Girl Baby Bodysuit Red Blue **Cranberry** Funny Cotton Baby Clothes"*

Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions
- **Previous work:** deep neural networks for **sequence tagging**
- **Limitations of previous work:**

(-) designed for a single category

(-) ignore product **categories**

(-) hard to capture **diversity** of categories

Digital Camera

flavor?
Not applicable

Vitamin

flavor: “fruit”

Fruit

flavor: ~~“fruit”~~
Not valid

Attribute Value Extraction from Product Profiles

- **Goal:** extract attribute values from product titles & descriptions
- **Previous work:** deep neural networks for **sequence tagging**
- **Limitations of previous work:**

(-) designed for a single category

(-) ignore product **categories**

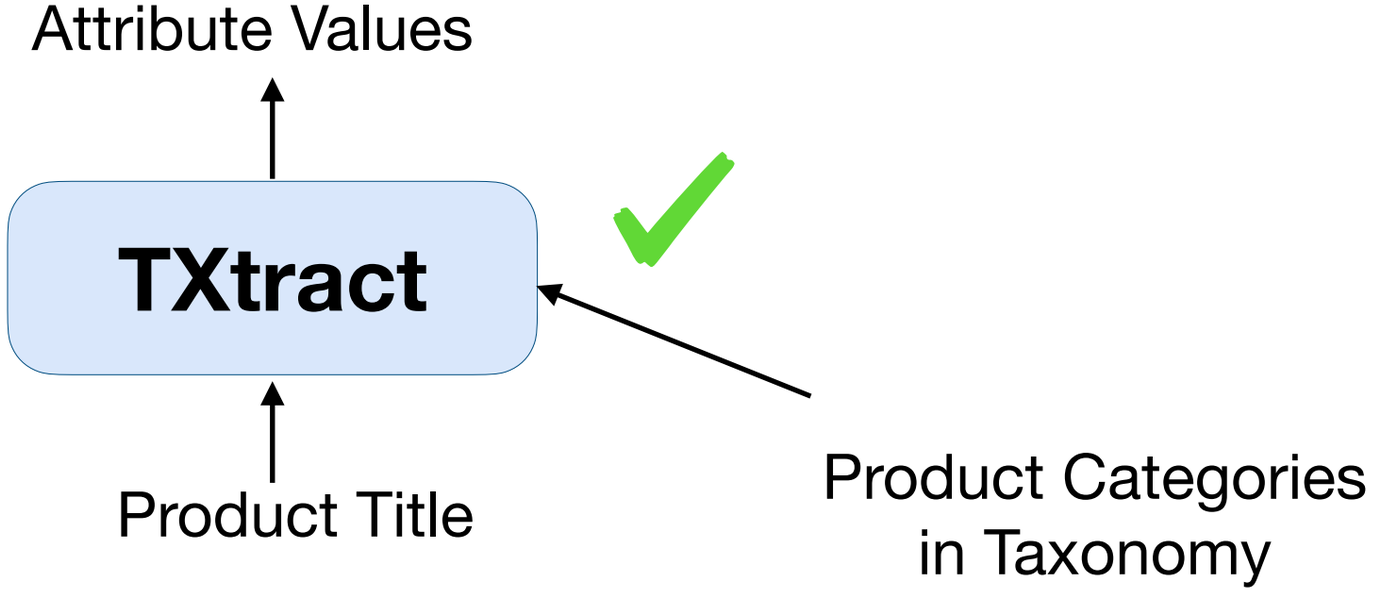
(-) hard to capture **diversity** of categories

(-) hard to scale to **large** product taxonomies in e-Commerce

- **>100M products**
- **>10K categories**
- **Products/categories continuously added**

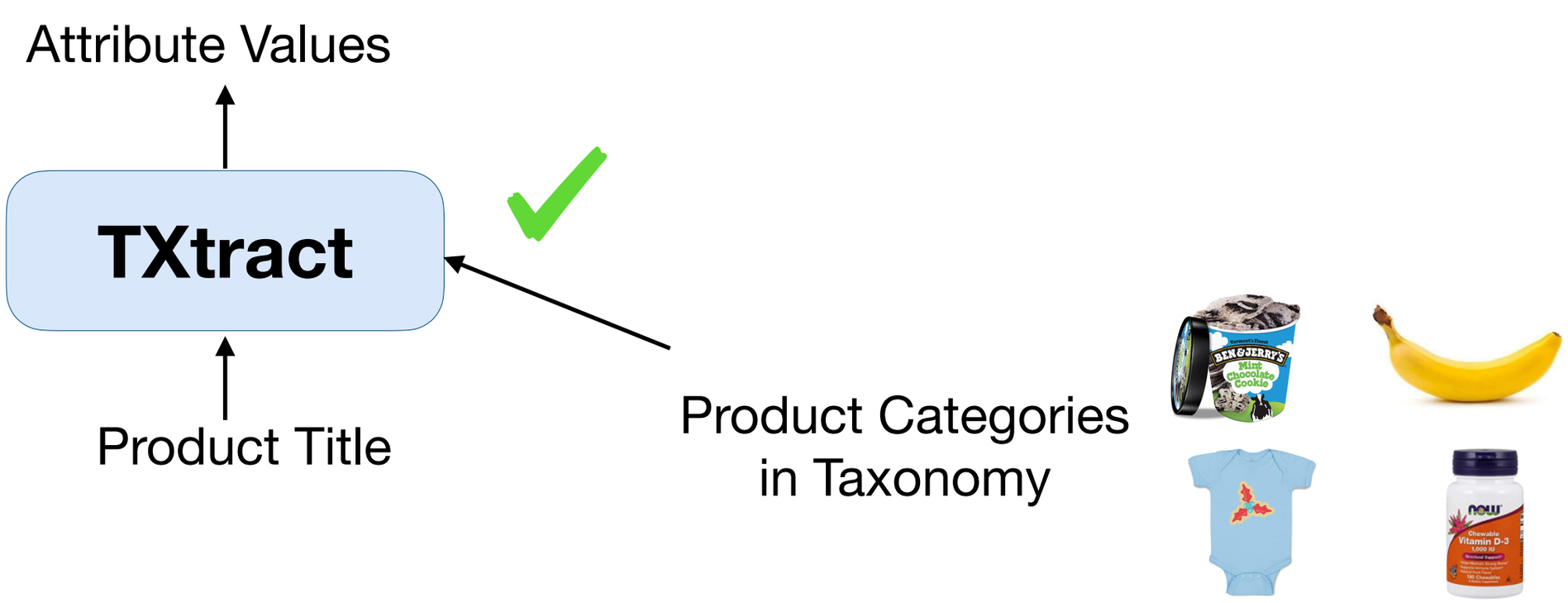
TXtract: Extraction for Thousands of Product Categories

- **TXtract:** a taxonomy-aware neural network for attribute value extraction



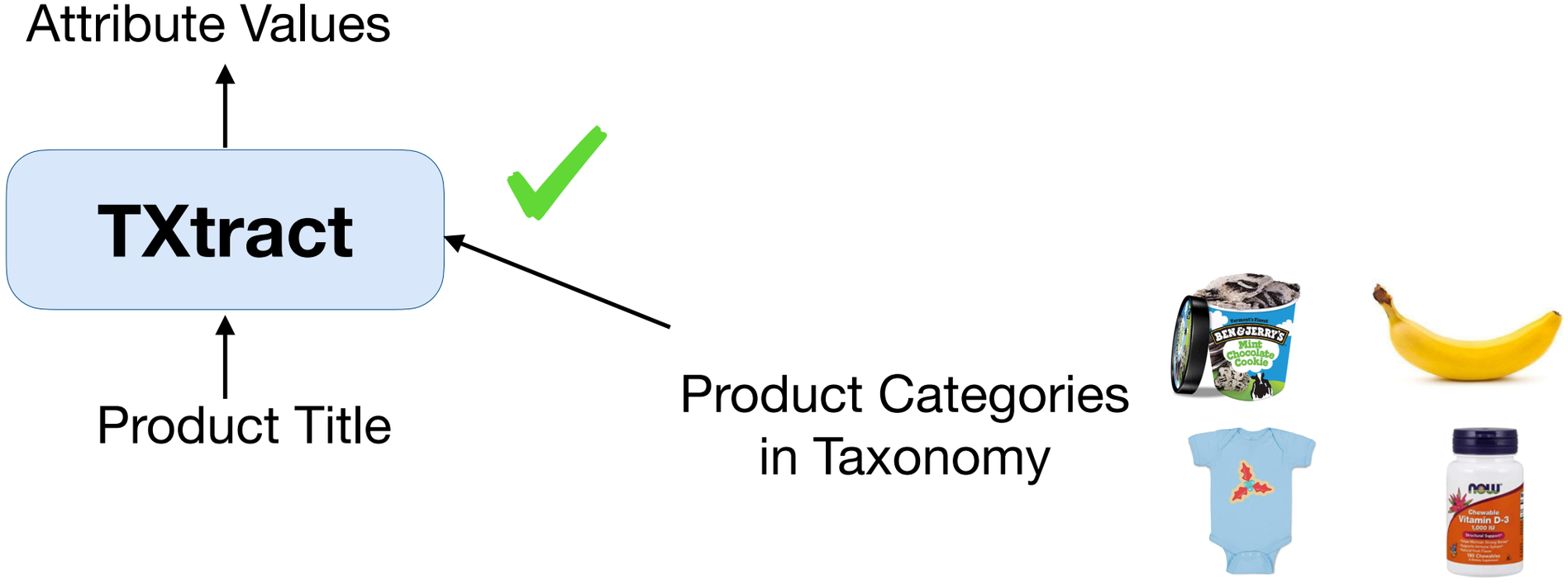
TXtract: Extraction for Thousands of Product Categories

- **TXtract:** a taxonomy-aware neural network for attribute value extraction
- **Our Contributions:**
  1. Consider **multiple** categories efficiently with a **single** model



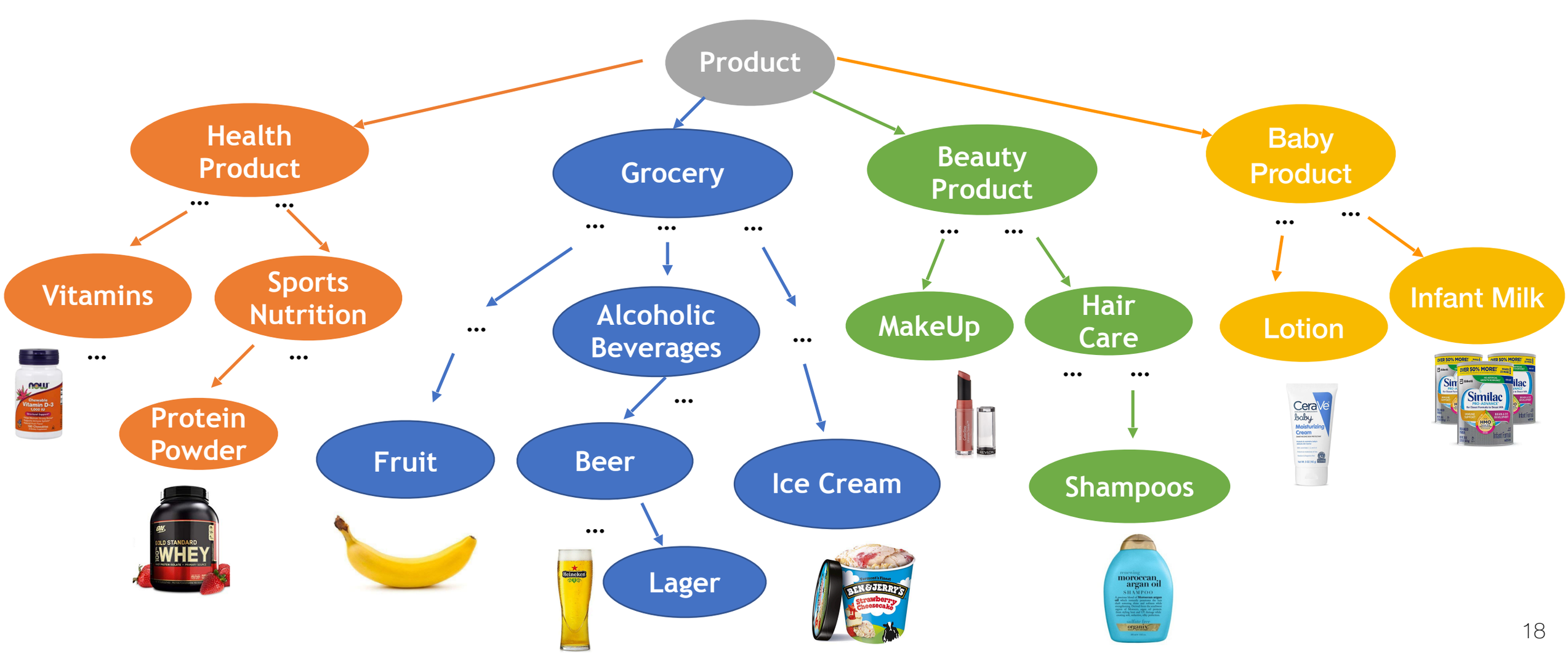
TXtract: Extraction for Thousands of Product Categories

- **TXtract:** a taxonomy-aware neural network for attribute value extraction
- **Our Contributions:**
  1. Consider **multiple** categories efficiently with a **single** model
  2. Extract **category-specific** attribute values using **conditional** self-attention



TXtract: Extraction for Thousands of Product Categories

- **TXtract:** a taxonomy-aware neural network for attribute value extraction
- **Our Contributions:**
  1. Consider **multiple** categories efficiently with a **single** model
  2. Extract **category-specific** attribute values using **conditional** self-attention
  3. **Scale up** extraction to hierarchical taxonomies with **thousands** of categories

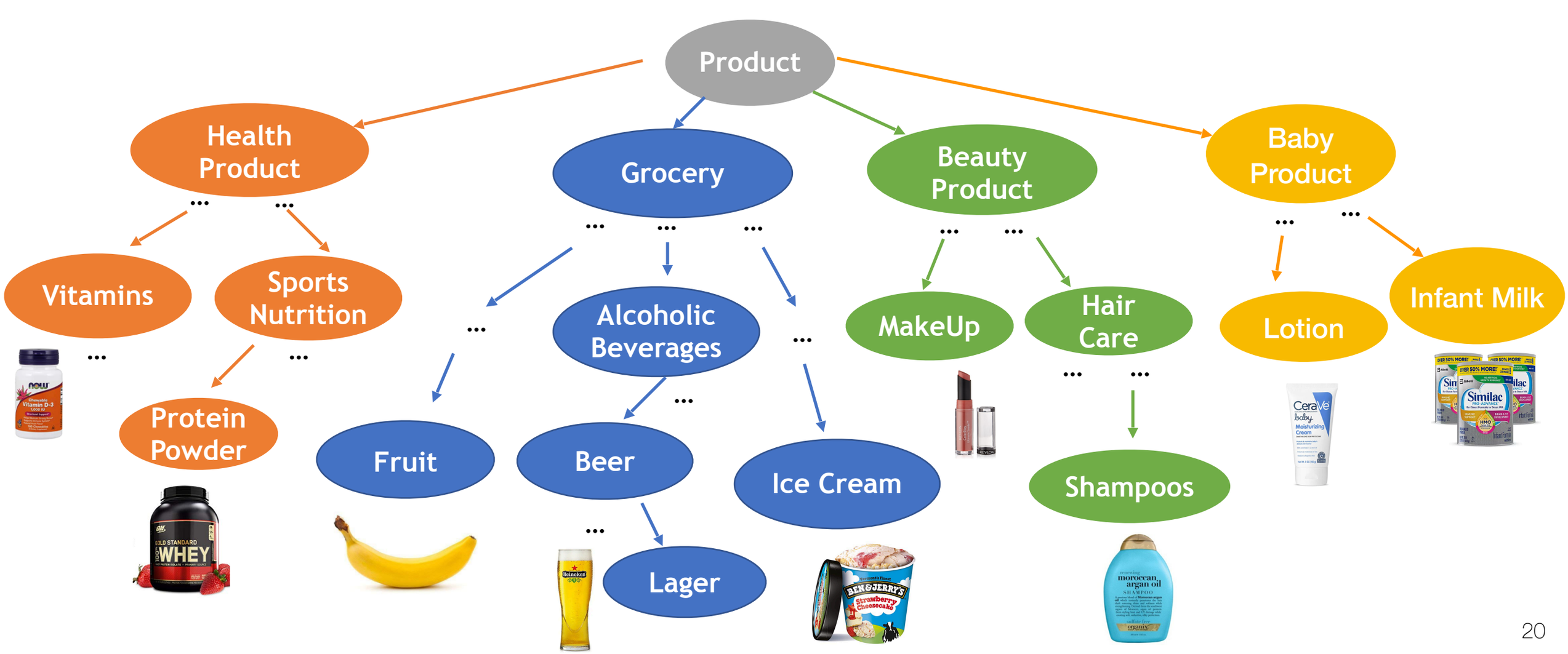


TXtract: Extraction for Thousands of Product Categories

- **TXtract:** a taxonomy-aware neural network for attribute value extraction
- **Our Contributions:**
  1. Consider **multiple** categories efficiently with a **single** model
  2. Extract **category-specific** attribute values using **conditional** self-attention
  3. **Scale up** extraction to hierarchical taxonomies with **thousands** of categories

TXtract: Extraction for Thousands of Product Categories

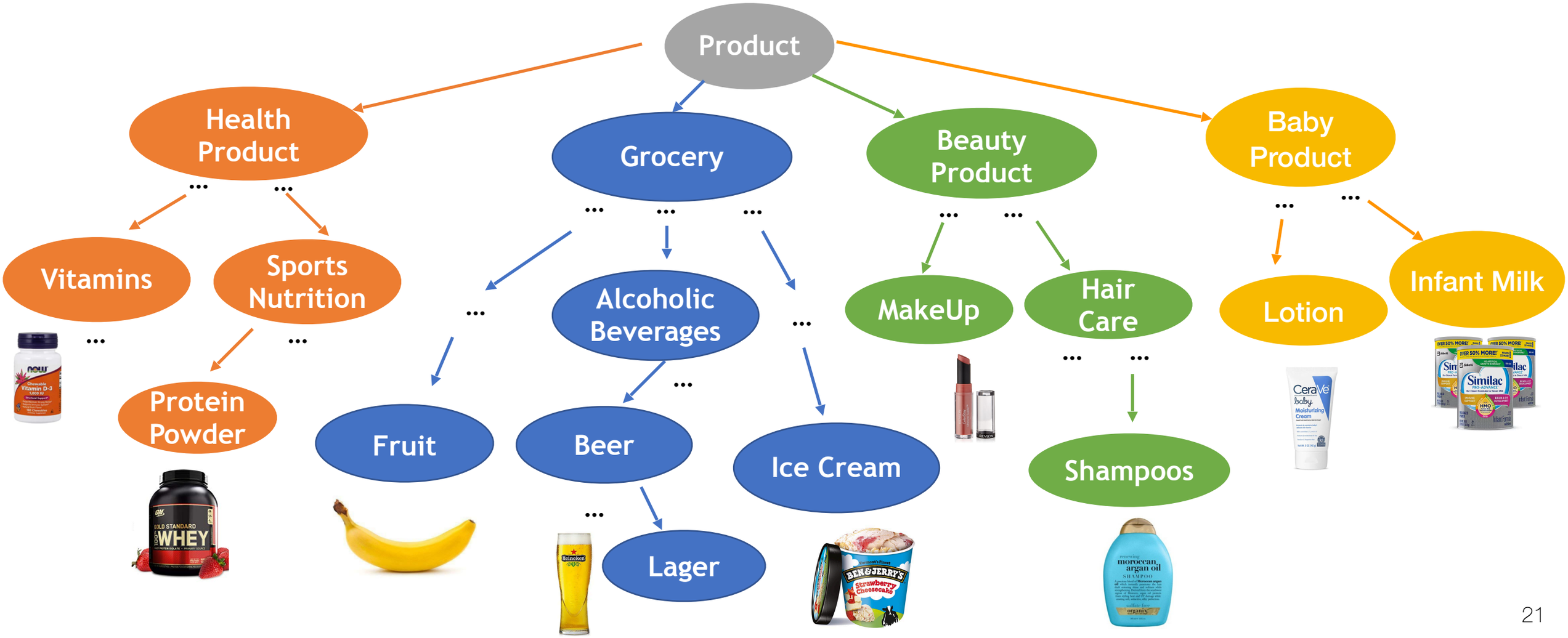
- **TXtract:** a taxonomy-aware neural network for attribute value extraction
- **Our Contributions:**
  1. Consider **multiple** categories efficiently with a **single** model
  2. Extract **category-specific** attribute values using **conditional** self-attention
  3. **Scale up** extraction to hierarchical taxonomies with **thousands** of categories



TXtract: Extraction for Thousands of Product Categories

- **TXtract:** a taxonomy-aware neural network for attribute value extraction

- **Our Contributions:**
  1. Consider **multiple** categories efficiently with a **single** model
  2. Extract **category-specific** attribute values using **conditional** self-attention
  3. **Scale up** extraction to hierarchical taxonomies with **thousands** of categories
  4. Increase **robustness** to wrong category assignments using **multi-task** training



Outline

1. Attribute Value Extraction from Product Profiles

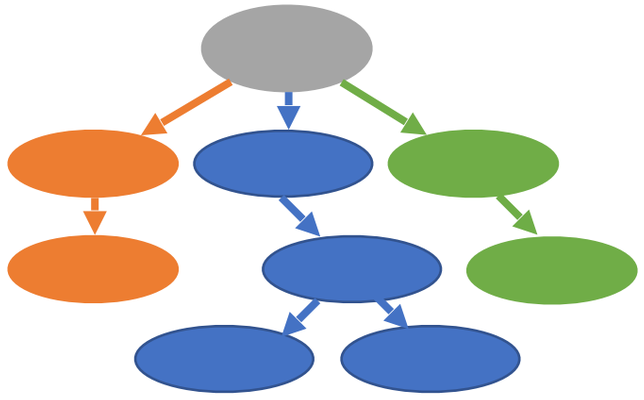
2. TXtract: Taxonomy-Aware Attribute Value Extraction

3. Experiments

4. Conclusions and Ongoing Work

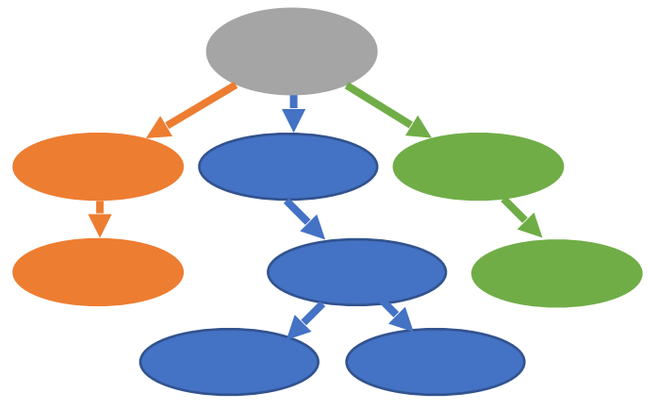
Scaling to Thousands of Product Categories - Challenges

- **Goal:**
  - Extracts attribute values for products …
  - … from thousands of **diverse** categories
  - … organized in **hierarchical** taxonomies



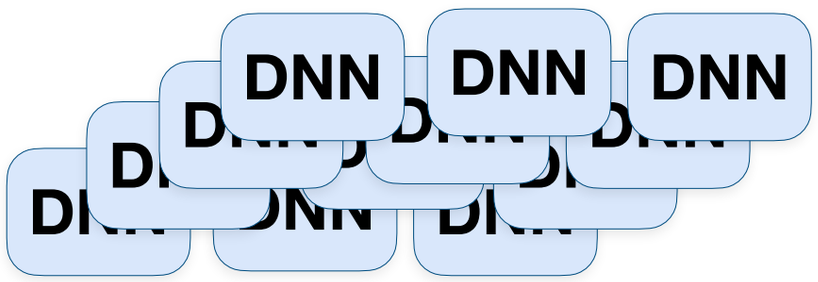
Scaling to Thousands of Product Categories - Challenges

- **Goal:**
  - Extracts attribute values for products …
  - … from thousands of **diverse** categories
  - … organized in **hierarchical** taxonomies



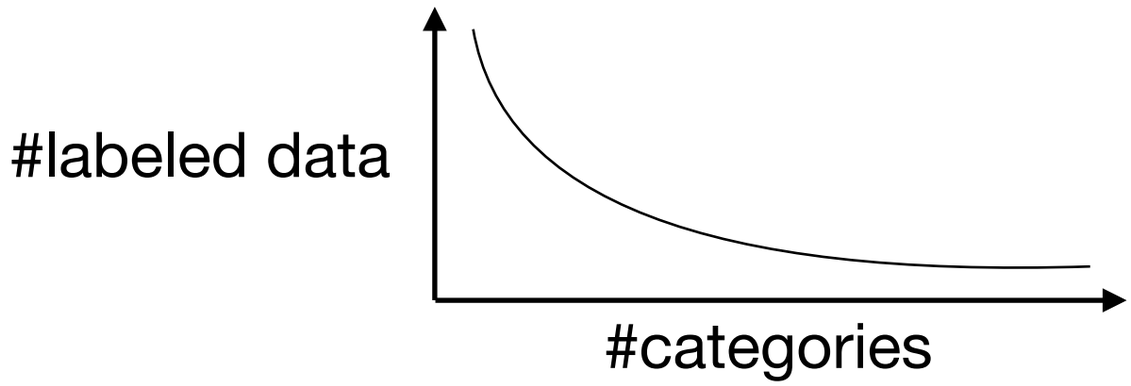
- **Approach1:** train a **separate** DNN for each category?

(-) expensive



store/orchestrate
1000+ models

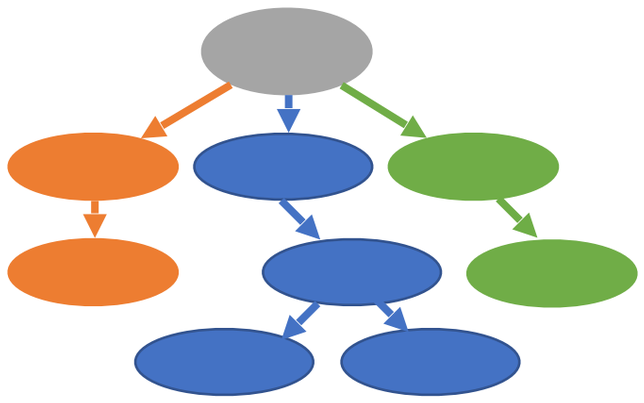
(-) prone to overfitting



most categories have
<<1000 labeled training data

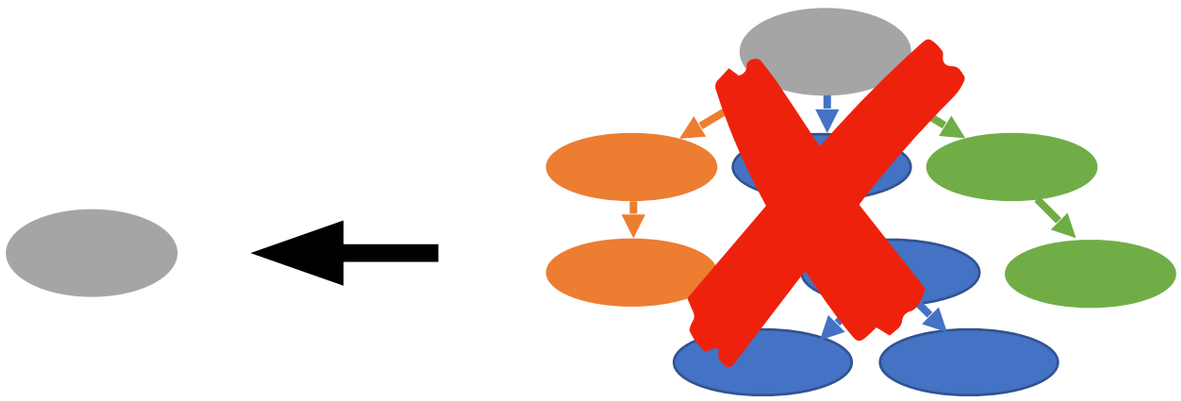
Scaling to Thousands of Product Categories - Challenges

- **Goal:**
  - Extracts attribute values for products …
  - … from thousands of **diverse** categories
  - … organized in **hierarchical** taxonomies



- **Approach1:** train a **separate** DNN for each category?

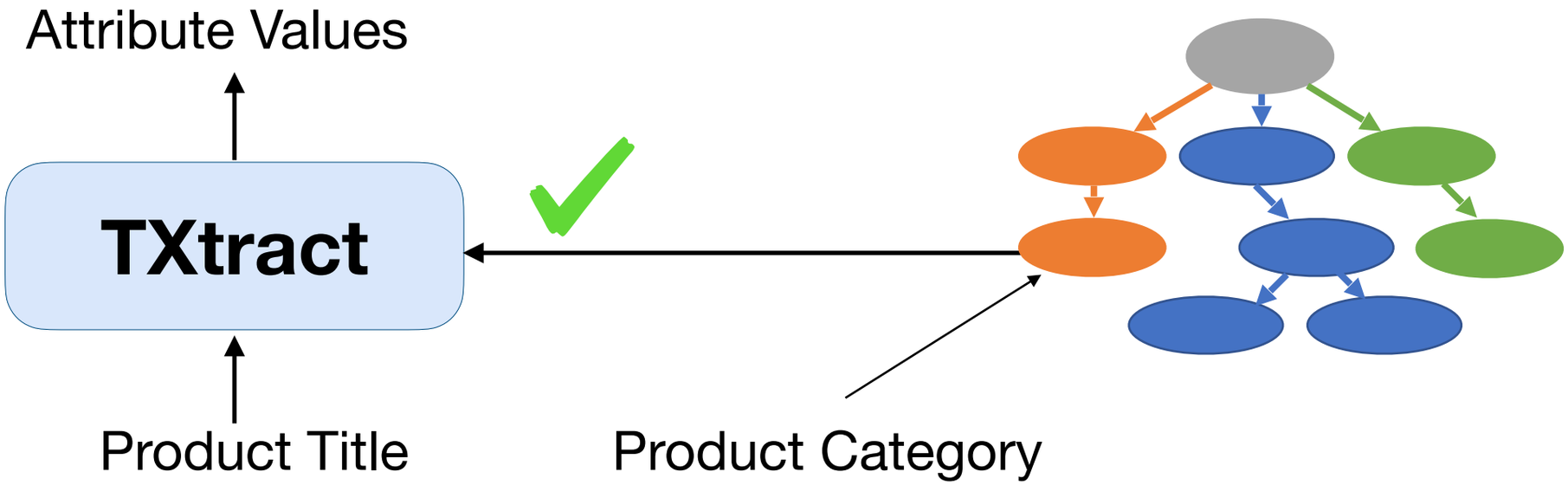
- **Approach 2:** assume a single “flat” category?



(-) not effective: missing category-specific characteristics

TXtract: Taxonomy-Aware Attribute Value Extraction

- TXtract leverages the **hierarchical** product taxonomy

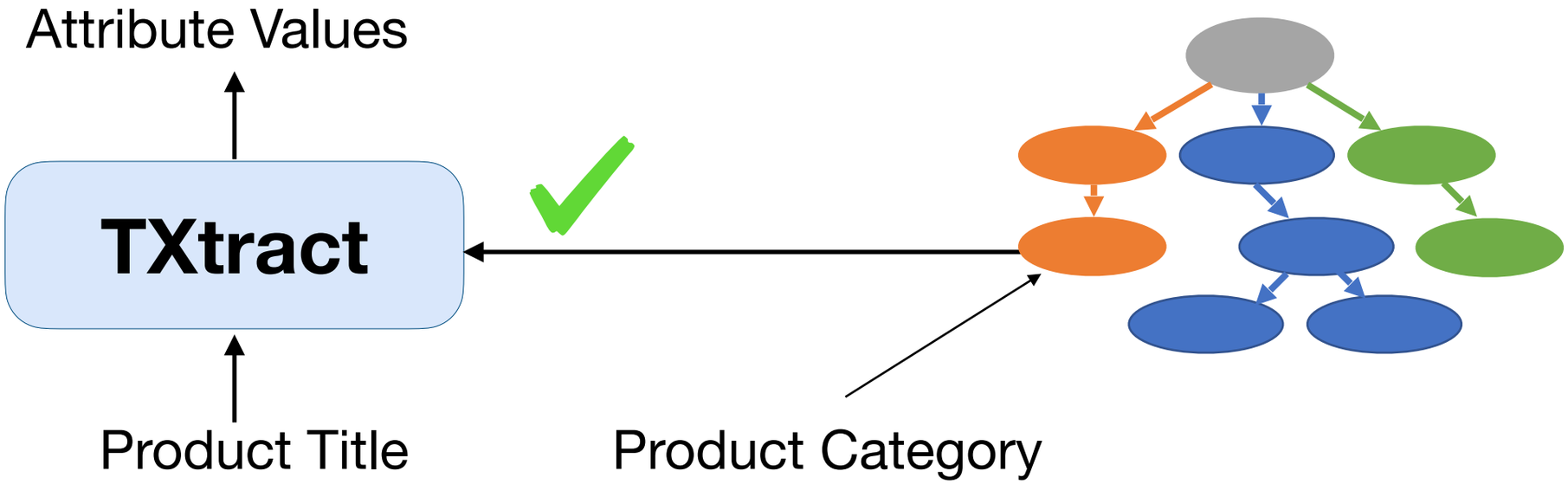


(+) efficient: single model for **all** categories

- "Small" categories: leverage products from **related** categories

TXtract: Taxonomy-Aware Attribute Value Extraction

- TXtract leverages the **hierarchical** product taxonomy



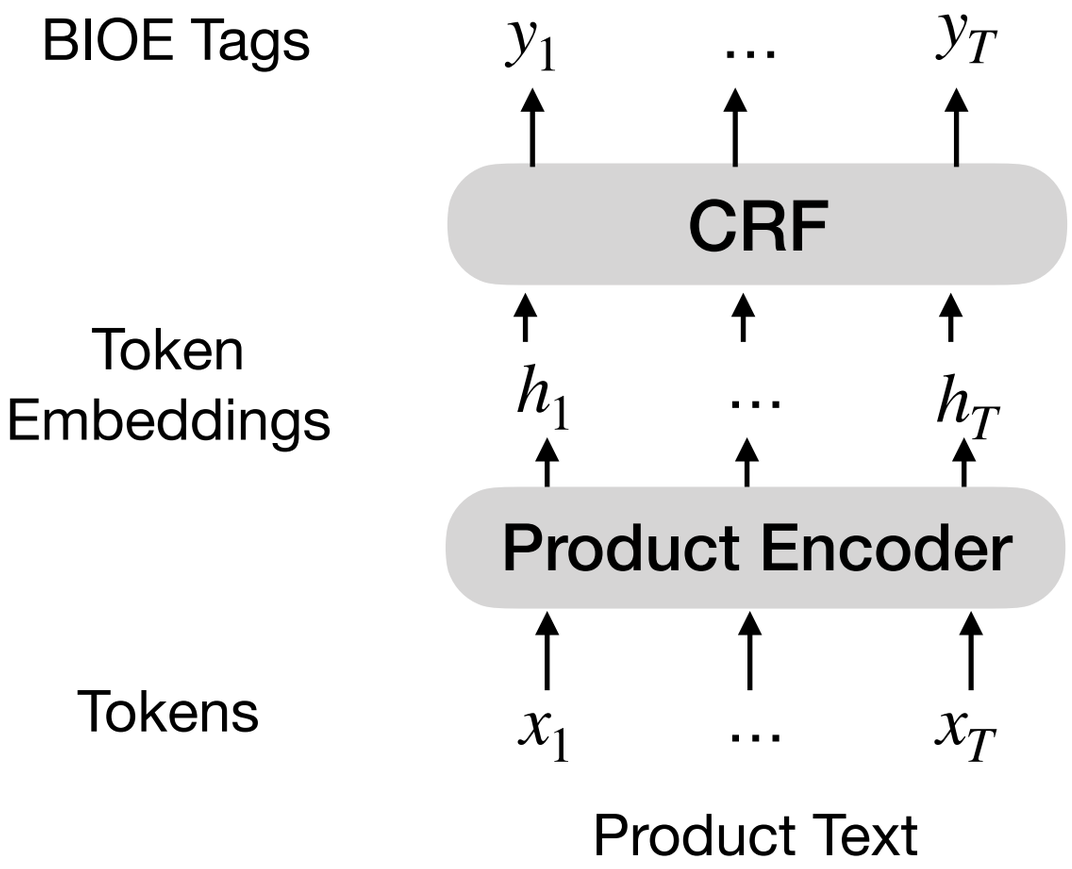
(+) efficient: single model for **all** categories

(+) effective: extracts **category-specific** attribute values

- product category -> attribute applicability, valid attribute values

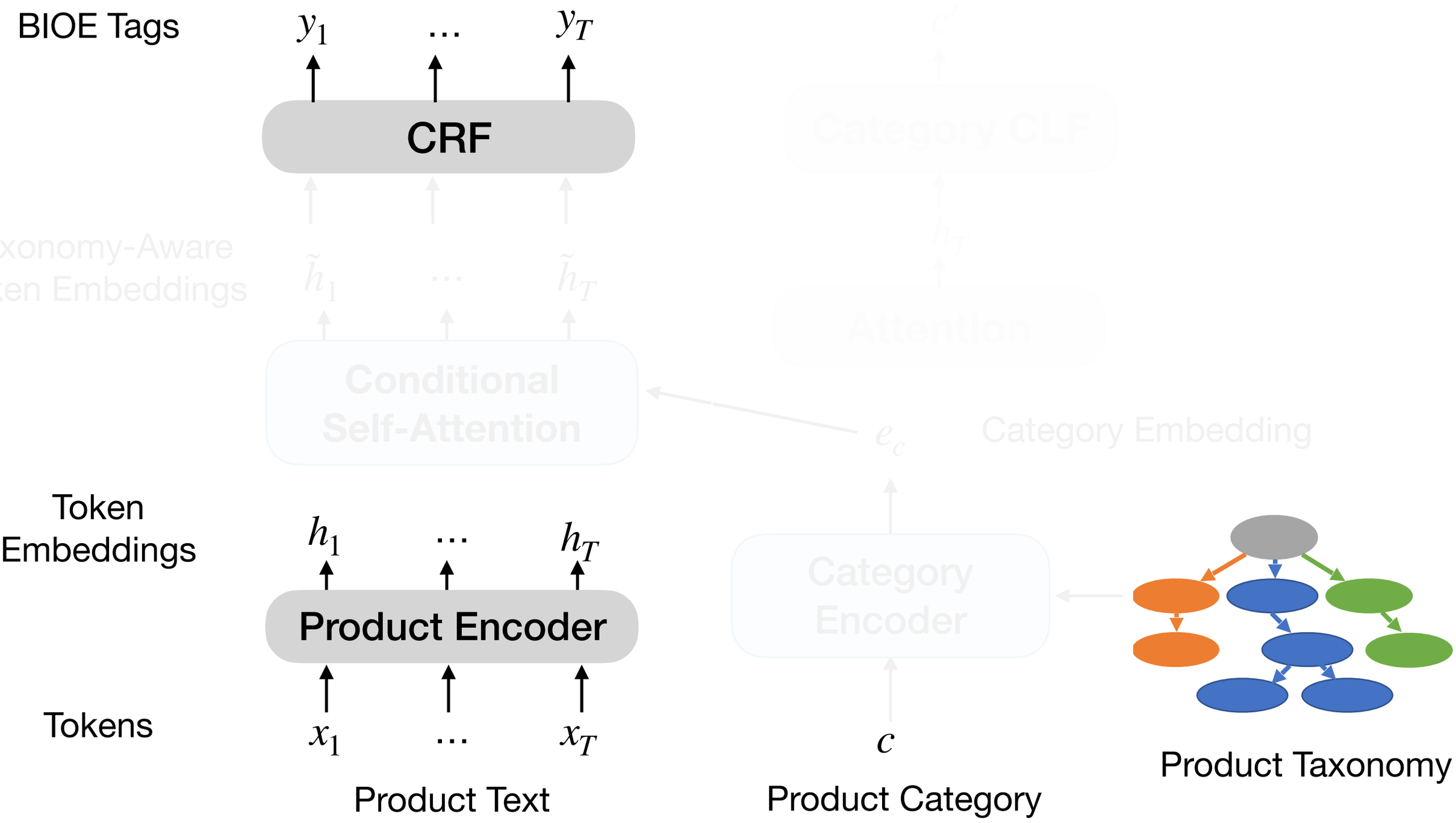
Leveraging Hierarchical Product Categories in TXtract

previous models

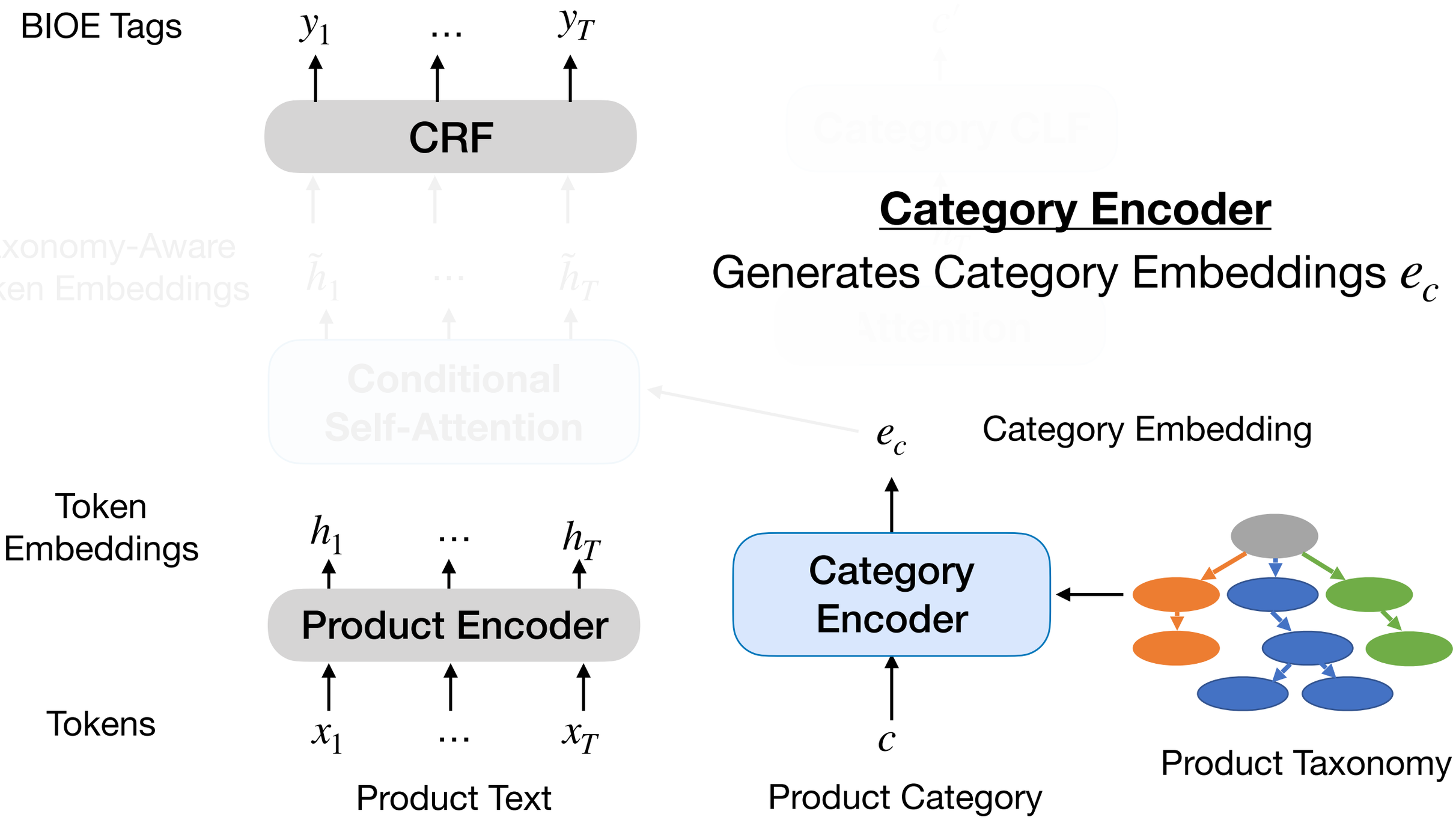


CRF: Conditional Random Field

Leveraging Hierarchical Product Categories in TXtract



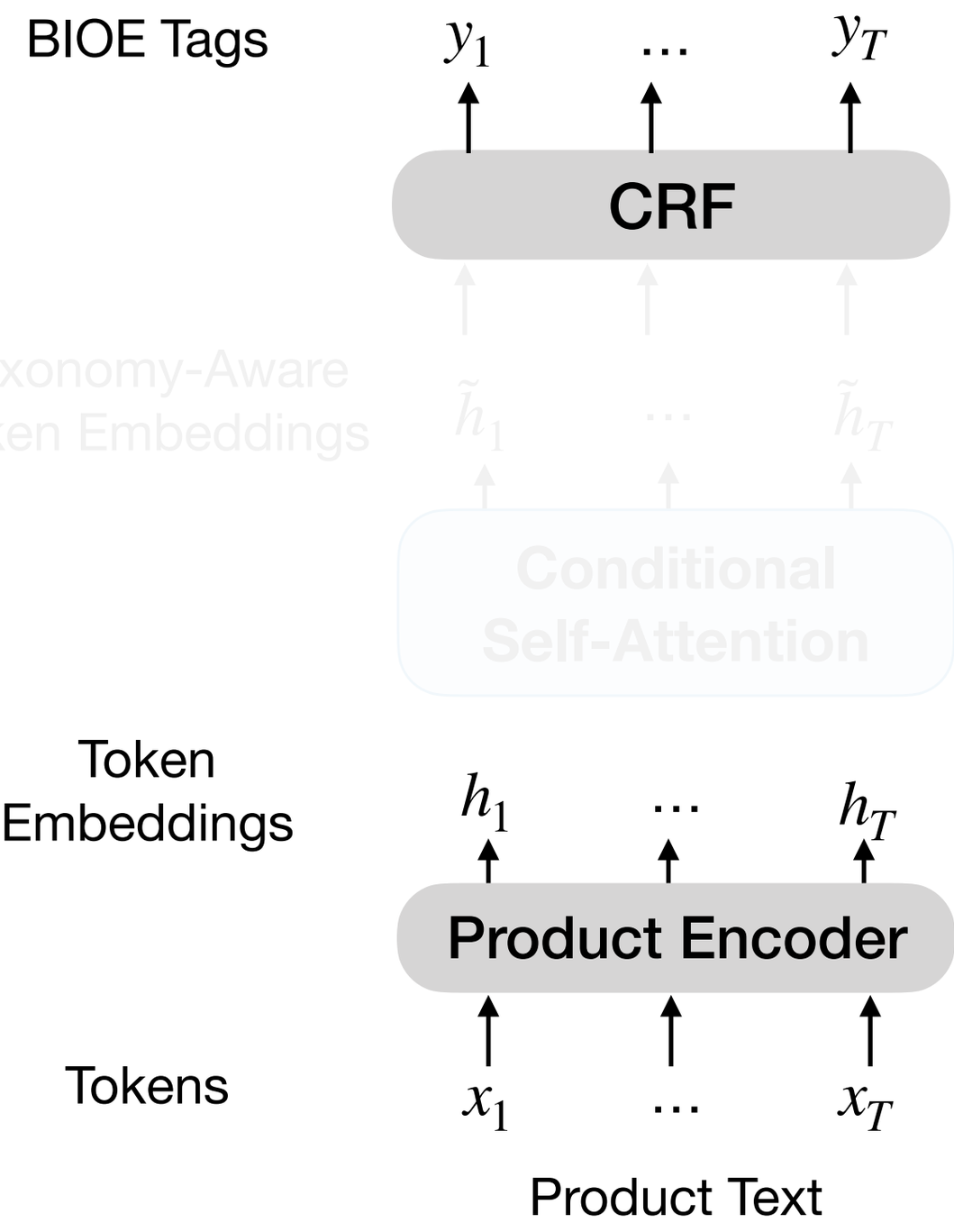
Leveraging Hierarchical Product Categories in TXtract



Category Encoder

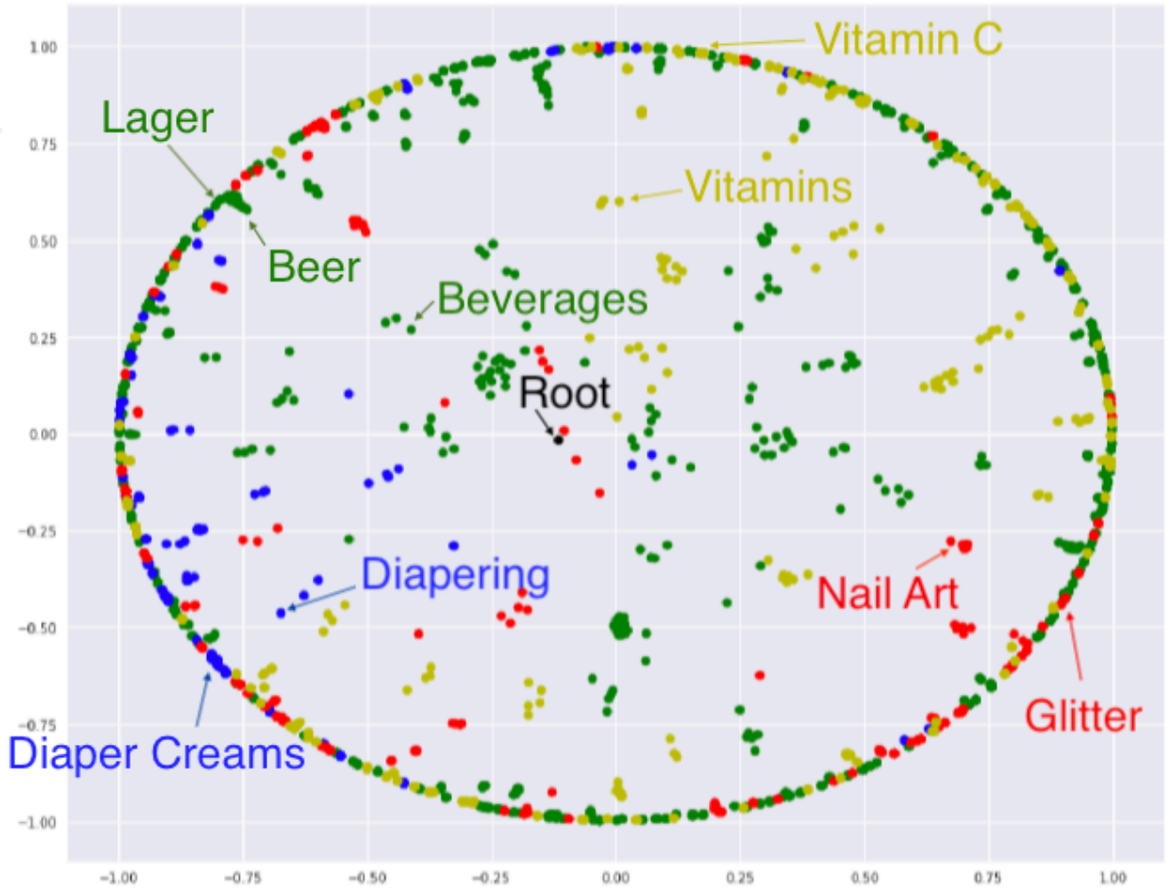
Generates Category Embeddings e_c

Leveraging Hierarchical Product Categories in TXtract

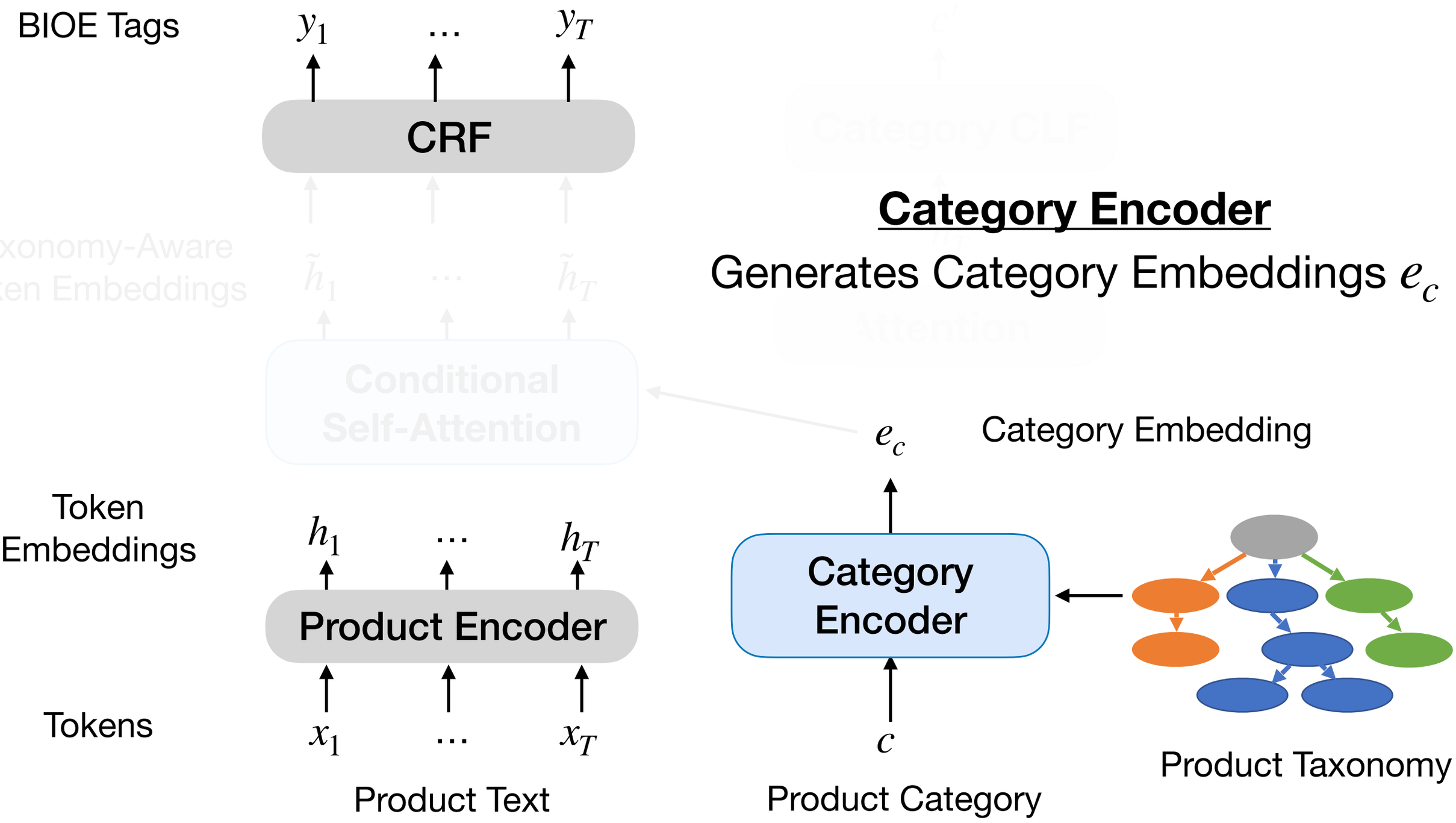


Category Encoder

Generates Category Embeddings e_c



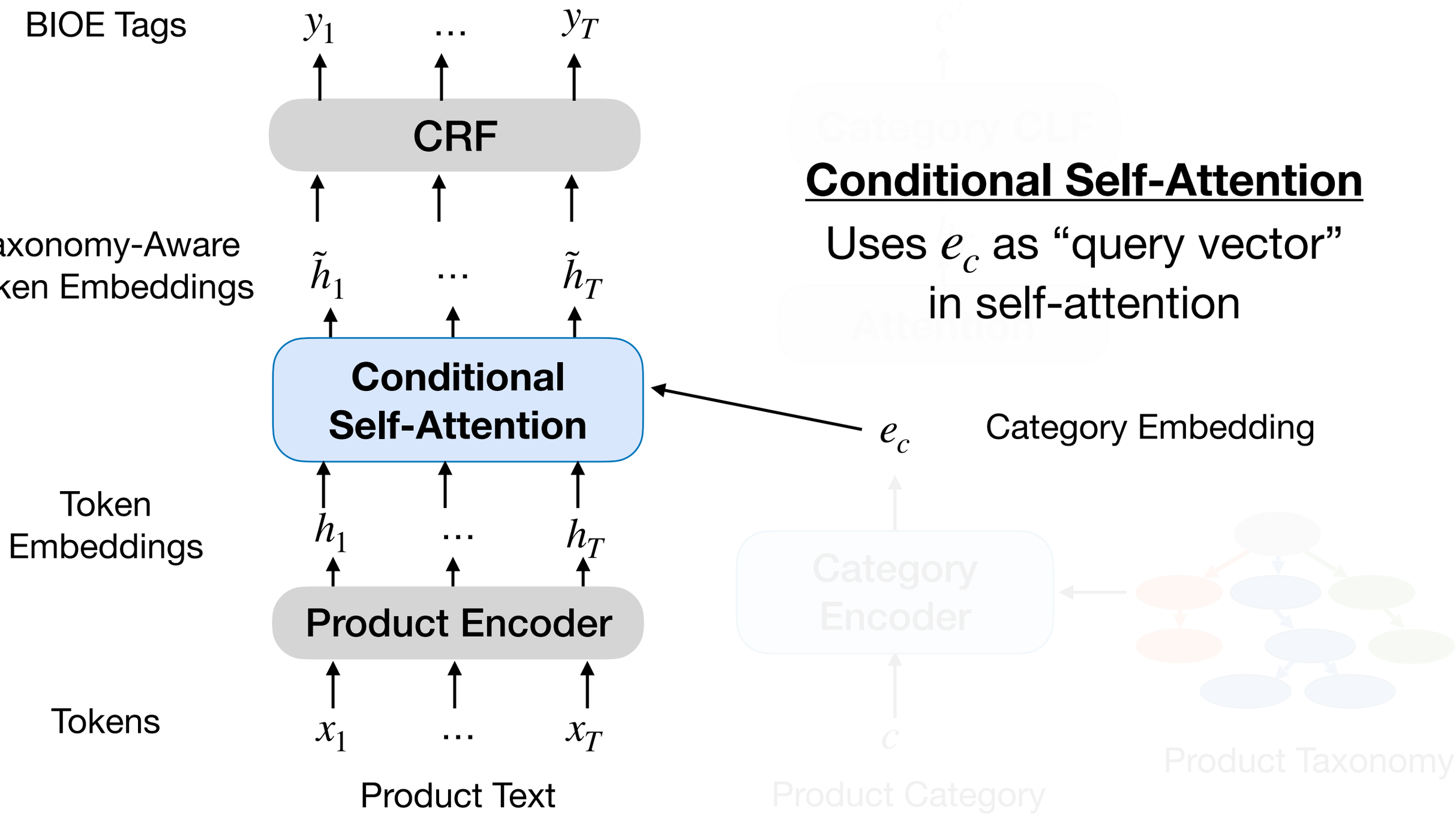
Leveraging Hierarchical Product Categories in TXtract



Category Encoder

Generates Category Embeddings e_c

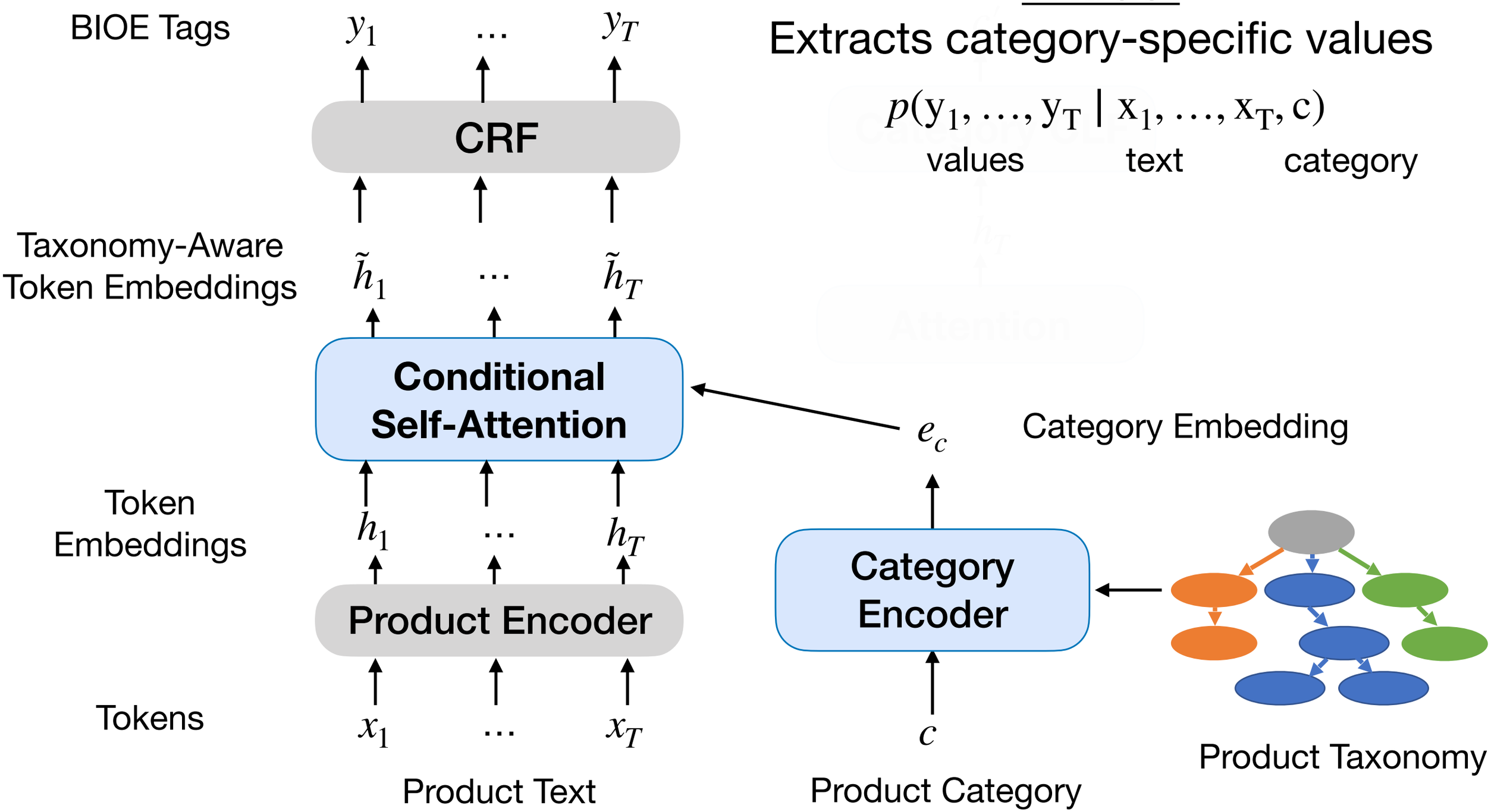
Leveraging Hierarchical Product Categories in TXtract



Conditional Self-Attention

Uses e_c as "query vector" in self-attention

Leveraging Hierarchical Product Categories in TXtract



TXtract

Extracts category-specific values

$$p(y_1, \ldots, y_T \mid x_1, \ldots, x_T, c)$$

values text category

Improving Robustness Towards Wrong Category Assignments

TXtract

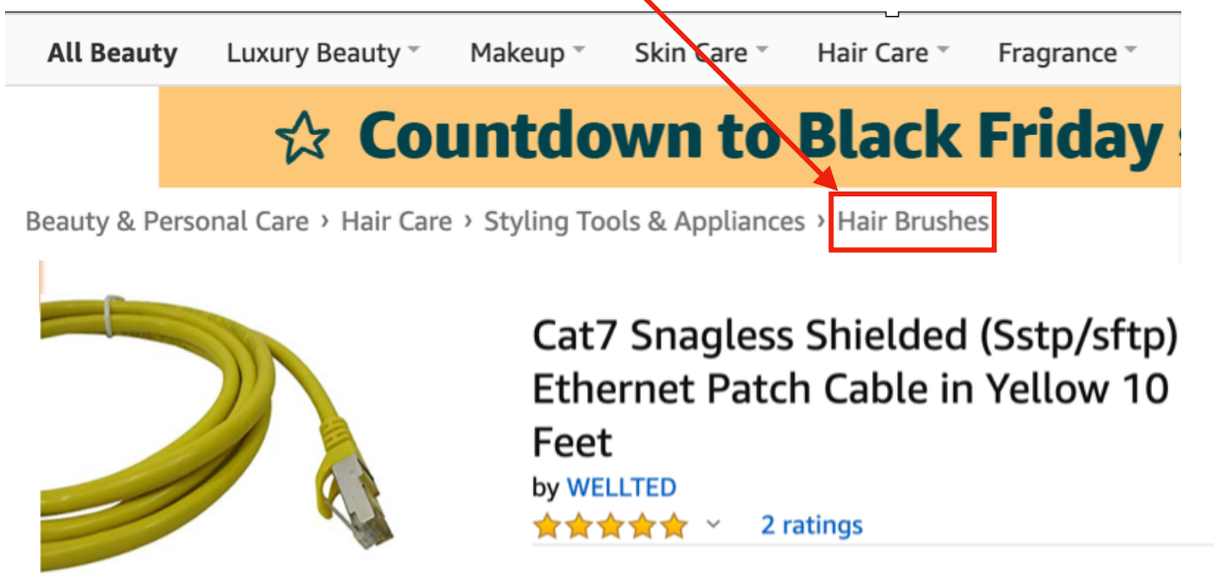
Extracts category-specific values

$$p(y_1, \ldots, y_T \mid x_1, \ldots, x_T, c)$$

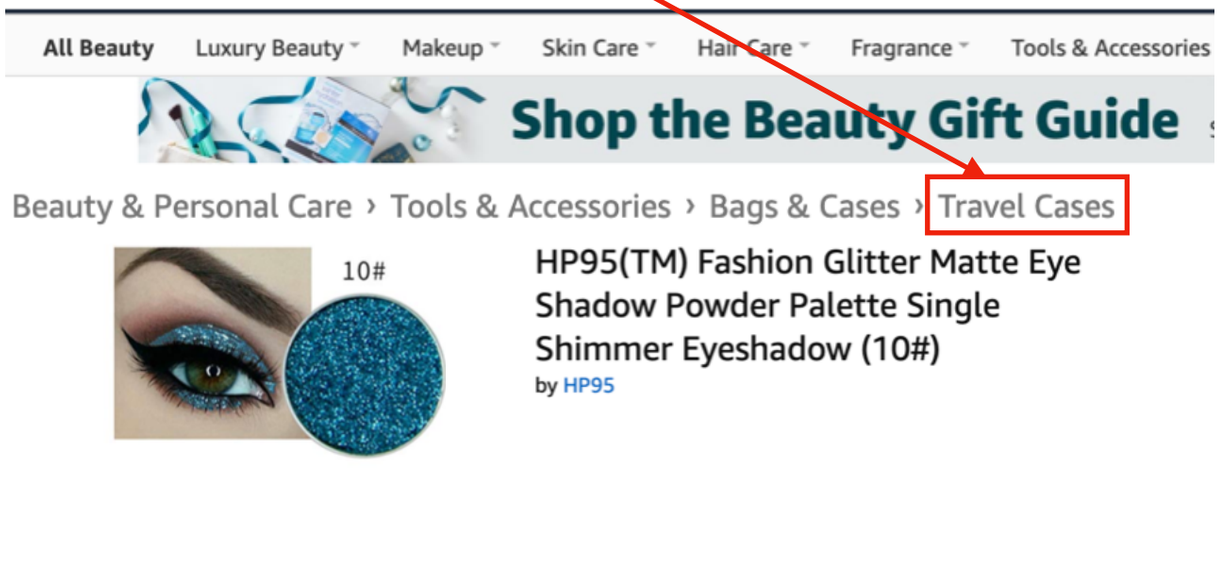
values text category

(-) Issue: products may be assigned to **wrong** taxonomy nodes!

Ethernet cable assigned under
Hair Brushes



Eyeshadow assigned under
Travel Cases



Improving Robustness Towards Wrong Category Assignments

TXtract

Extracts category-specific values

$$p(y_1, \ldots, y_T \mid x_1, \ldots, x_T, c)$$

values ***wrong*** | text | category ***wrong***

(-) Issue: products may be assigned to **wrong** taxonomy nodes!

(-) Conditioning on **wrong** categories -> **wrong** values

Improving Robustness Towards Wrong Category Assignments

Main Task

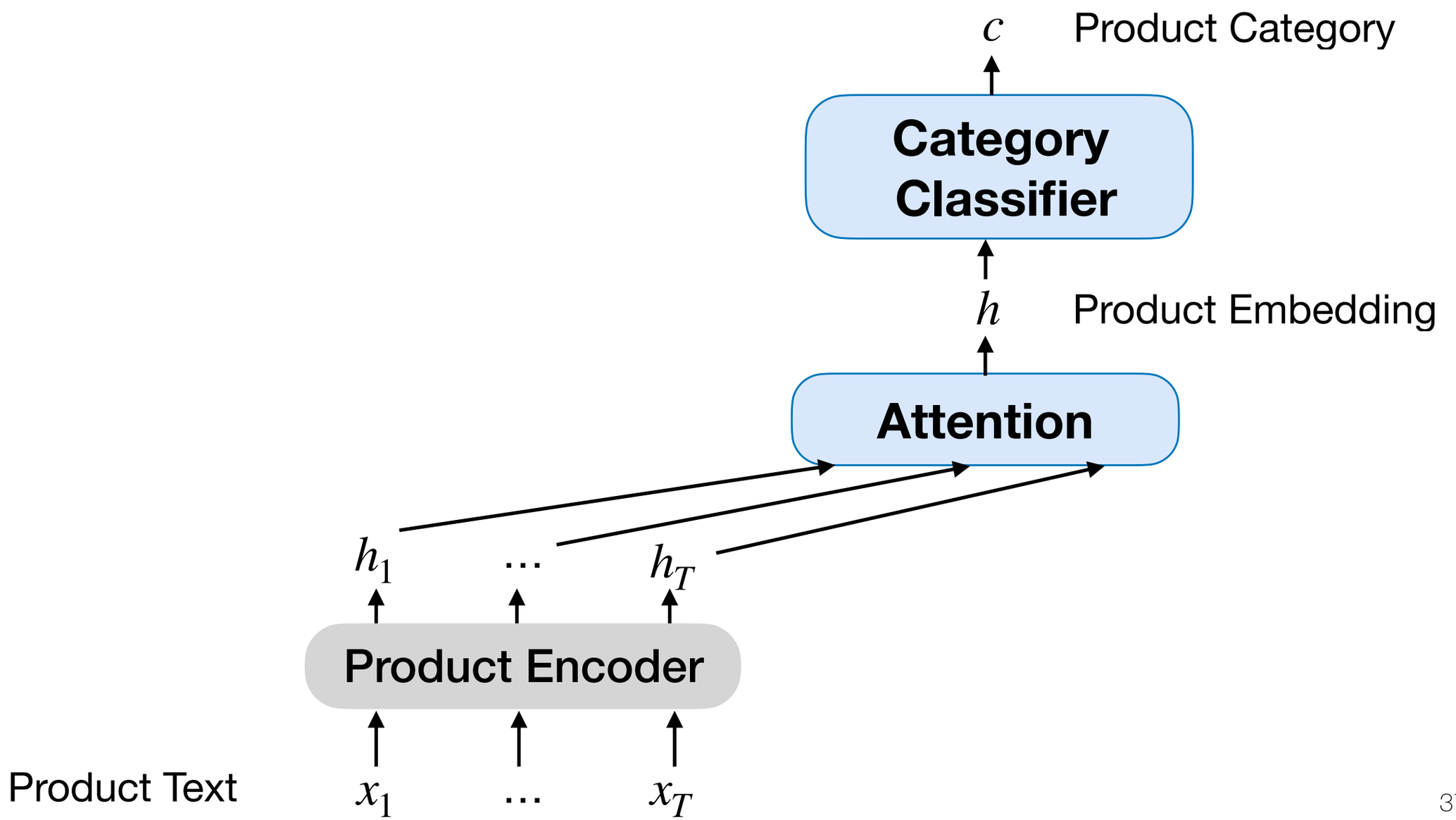
Extract category-specific values

$$p(y_1, \ldots, y_T \mid x_1, \ldots, x_T, c)$$

Auxiliary Task

Predict categories from text

$$p(c \mid x_1, \ldots, x_T)$$



Improving Robustness Towards Wrong Category Assignments

Main Task

Extract category-specific values

$$p(y_1, \ldots, y_T \mid x_1, \ldots, x_T, c)$$

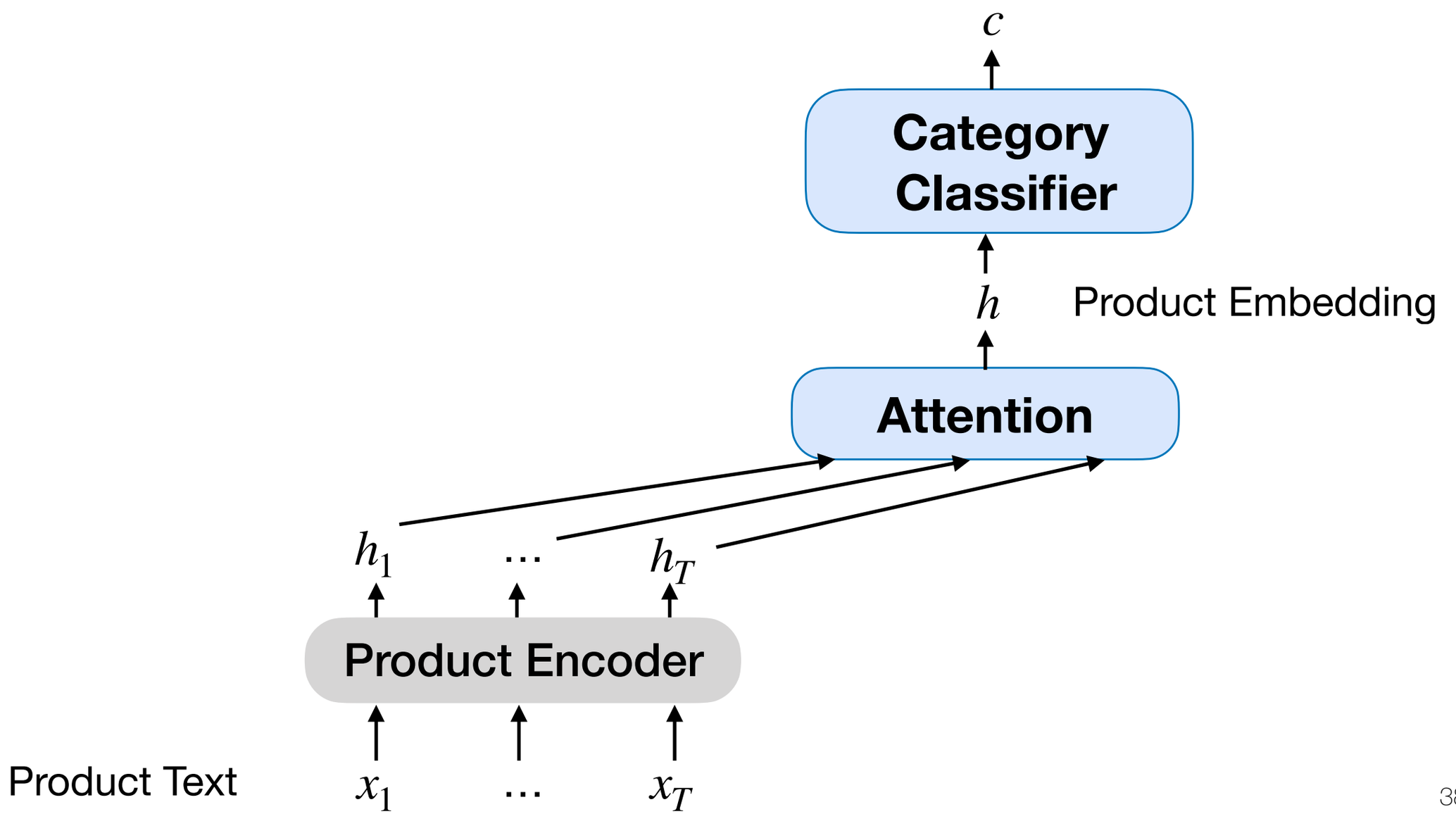
Auxiliary Task

Predict categories from text

$$p(c \mid x_1, \ldots, x_T)$$

"Taxonomy-aware" loss function

"Correctly guess category AND ancestors"



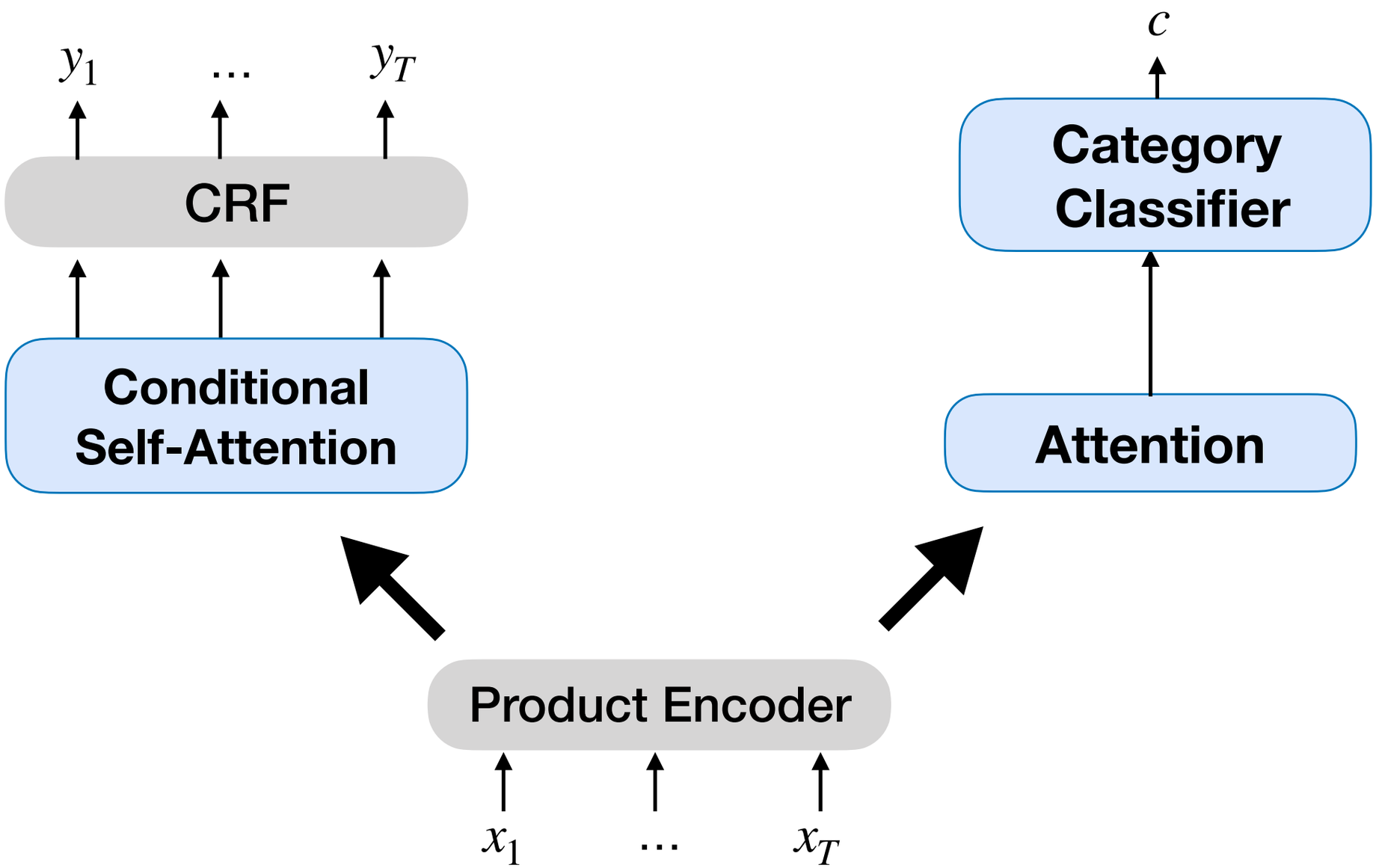
Multi-Task Training of TXtract

Main Task

Taxonomy-Aware
Attribute Value Extraction

Auxiliary Task

Taxonomy-Aware
Category Prediction



Outline

1. Attribute Value Extraction from Product Profiles

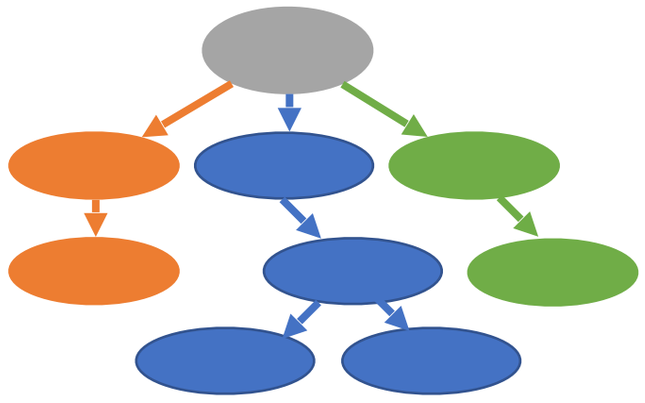
2. TXtract: Taxonomy-Aware Attribute Value Extraction

3. Experiments: Taxonomy with 4,000 Product Categories

4. Conclusions and Ongoing Work

Experiments: Attribute Value Extraction

- **Dataset:**
  - 2 million products *(sampled from Amazon.com webpages)*
  - 4,000 categories *(sampled from Amazon's taxonomy)*



- **Attributes:** *brand*, *flavor*, *package size, ingredients*
- **Training:** distant supervision for sequence tagging

Catalog values

Product ID	Brand	Flavor	Size	Ingredients
B00FZHEGGW	Fage	Plain	35.3 oz	...
B0725VRRLP	Ben & Jerry's	...	...	...
...	...	...	...	...

flavor tags

Input	Ben	&	Jerry's	black	cherry	cheesecake	ice	cream
Output	O	O	O	B	I	E	O	O

TXtract Effectively Leverages Product Categories

Average performance
across **ALL** categories & attributes

[Zheng et al., KDD'18]

ignores categories →

considers categories →

	Coverage (%)	**Macro F1 (%)**
OpenTag	73.0	46.6
TXtract	**81.6** (**+11.7%**)	**49.7** (**+10.4%**)

- TXtract outperforms OpenTag across 4,000 categories

TXtract Effectively Leverages Product Categories

Average performance
across **ALL** categories & attributes

[Zheng et al., KDD'18]
ignores categories →

considers categories →

	Coverage (%)	**Macro F1 (%)**
OpenTag	73.0	46.6
TXtract	**81.6** (**+11.7%**)	**49.7** (**+10.4%**)

- TXtract outperforms OpenTag across 4,000 categories
- TXtract outperforms other **category-aware** approaches [Cho et al., EMNLP'14] [Johnson et al., TACL'17] [Ma et al., KDD'19]

See more results and ablation study in our paper!

Outline

1. Attribute Value Extraction from Product Profiles

2. TXtract: Taxonomy-Aware Attribute Value Extraction

3. Experiments: Taxonomy with 4,000 Product Categories

4. Conclusions and Ongoing Work

Attribute Value Extraction - Scaling Up to Thousands of Product Categories

- E-commerce domain is challenging!
  - Diverse categories

Digital Camera

flavor?

Not applicable

Vitamin

flavor: "fruit"

Fruit

flavor: ~~"fruit"~~

Not valid

  - Assignments to wrong categories

Hair Brush

Attribute Value Extraction - Scaling Up to Thousands of Product Categories

- E-commerce domain is challenging!
  - Diverse categories

  - Assignments to wrong categories

- TXtract: **hierarchical** taxonomies with **thousands** of categories
  - **(+) Efficient:**
    - single model trained on all categories in parallel
  - **(+) Effective:**
    - Leverages taxonomy using conditional self-attention & multi-task training
    - Improves extraction quality (e.g., up to 15% higher coverage)

Towards Better, Large-Scale Product Understanding

TXtract

flavor: “chocolate”

ingredients: “argan oil”, …

Towards Better, Large-Scale Product Understanding

Building an “automatic” knowledge graph of products

[Saldana et al., KDD’20]

TXtract

flavor: “chocolate”

ingredients: “argan oil”, …

[Saldana et al. KDD’20] *AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types*

Towards Better, Large-Scale Product Understanding

[Saldana et al., KDD'20]

TXtract

flavor: "chocolate"

ingredients: "argan oil", …

product graph

"Alexa, which shampoos contain argan oil?"

[Saldana et al. KDD'20] *AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types*

Thank you!

Giannis Karamanolakis
Columbia University
gkaraman@cs.columbia.edu
https://gkaramanolakis.github.io

Jun Ma, Xin Luna Dong
Amazon.com
{junmaa, lunadong}@amazon.com