QueryNER:
Segmentation of E-Commerce Queries

Chester Palen-Michel, Lizzie Liang, Zhe Wu,
Constantine Lignos

Motivation

- Queries are often treated as an unstructured sequence of tokens:
  vintage laura ashley romper girls
- **Chunking** allows us to find the boundaries between spans of tokens and identify the purpose of each span:
  [vintage] [laura ashley] [romper] [girls]
- This allows us to better determine:
  - Which portions of the query are most important to relevance?
  - Which portions of the query might be most safely dropped?
  - Can we weight different spans for relevance rather than tokens?
  - Can we link spans to a knowledge graph (e.g. known brands)?

Prior Work

- Focus on aspect-value extraction from titles or descriptions
- Sparse
  - Only small percentage of tokens are part of a entity span
- Limited publicly available datasets

Aspect-Value Extraction:
High - end [speaker cover] for [B & W] [805d] 1 pair made of [velvet] [suede] made to order
QueryNER Segmentation:
[High - end] [speaker cover] for [B & W] [805d] [1 pair] [made of velvet] [suede] [made to order]

Ontology Design

- 17 entity types
- Full set of guidelines in appendix
- Manageable number of types for annotators
- Coverage of as much of the query tokens as possible

Entity Type	Count
Core Product Type	8,310
Modifier	3,367
Creator	2,217
Department	1,652
Product Name	1,345
Content	1,301
UoM	862
Color	691
Shape	607
Material	569
Occasion	397
Condition	178
Quantity	104
Price	51
Origin	40
Time	32
Product Number	31

Ontology Examples

Core Product Type: The main thing being sold. Generic ways of describing a product. These are not official product names but common objects.

Creator: The company or person who creates or produces the product. It could also be the designer name associated with the product or brand name.

Content: Names of characters, titles of tv or movies, sayings or phrases that appear on or within the product itself. Many mugs, t-shirts, figurines, or comic books have some form of content or characters associated with them.

[enesco Creator] [mickey Content] [ornament Core Product Type] [christmas in the air Content]

Ontology Examples

Product Name: The specific name of a product or model name.

Modifier: Modifier is used for spans that clarify the type of product. This can describe certain features a project has like “2 in 1” or “high performance”. Modifier can also be used for constraining the type of a product.

Condition: The condition of the product. This describes whether the product is new or old and can go into more detail about things such as whether a product includes its original tags.

Time: An expression of the date or time associated with the product that is not a unit of measurement.

[2011 Time] [ford Creator] [edge sel Product Name] [back up Modifier] [camera Core Product Type] [used Condition]

Ontology Examples

Department: Category of the population the item was made for.

Unit of Measurement (UOM): Any way of measuring size or other unit of measurement. This can include everything from clothing sizes, to lengths and widths, car engine sizes, battery capacity, amount of memory in a computer, lens sizes for cameras.

Material: The material or physical entity that makes up the item.

Color: Description of the color, pattern, appearance related to the surface appearance or 2-dimensional design of the product.

[Florida gators Content] [nike Creator] [mens Department] [xl UoM] [polyester Material] [orange Color] [tee shirts Core Product Type]

Annotation Process

- Internal rounds of annotation to develop ontology and guidelines
- Test set multiple annotators and adjudicated
- We release annotator level decisions on the test set
- Training set with single annotator
- ***Not*** crowd-sourced
- Human annotated
- Annotators were paid market rates

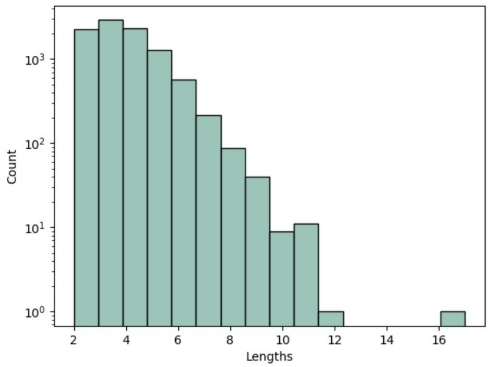
Brandeis

Dataset

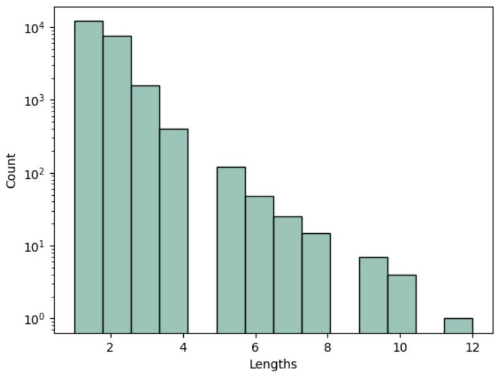
	Queries	Entities	Tokens
Train	7,841	17,505	28,457
Dev	871	1,930	3,124
Test	933	2,317	3,610

Average query length: 3.63 tokens
Average entity length: 1.6 tokens

Query length distribution



Entity length distribution



Experiments

	Precision	Recall	F1
BERT	$60.94_{\pm 0.5}$	$60.17_{\pm 0.4}$	$60.56_{\pm 0.4}$
XLM-R	$60.45_{\pm 0.5}$	$59.75_{\pm 0.5}$	$60.10_{\pm 0.5}$
BERT-cont.	$61.78_{\pm 0.4}$	$60.82_{\pm 0.3}$	$61.29_{\pm 0.3}$

Baselines with BERT and XLM-R

Continued masked language modeling on the rest of Amazon ESCI query data set provides a slight advantage

Brandeis

Experiments

Data Augmentation to create transformed test sets

Transformation	Example
Original	airforce 1 women shoes white
Shuffled	shoes women white airforce 1
Butterfinger	airvorce 1 women shoes white
Numeric	airforce 6 women shoes white
Color	airforce 1 women shoes green
Mention Replacement	zerogrand boys shoes leopard
All Transformations	shofs boys maple zerogrand

Training Data	Original		
Eval Data	Transformed Tests		
Augmentation	Precision	Recall	F1
butterfinger	$30.86_{\pm 0.6}$	$33.65_{\pm 0.5}$	$32.19_{\pm 0.5}$
numeric	$60.80_{\pm 0.7}$	$58.32_{\pm 0.5}$	$59.54_{\pm 0.6}$
mention replacement	$74.83_{\pm 1.1}$	$75.52_{\pm 1.0}$	$75.17_{\pm 1.1}$
color	$61.55_{\pm 0.6}$	$60.62_{\pm 0.4}$	$61.08_{\pm 0.5}$
shuffled	$59.82_{\pm 0.5}$	$59.41_{\pm 0.5}$	$59.62_{\pm 0.4}$
all transformations	$27.70_{\pm 0.4}$	$30.53_{\pm 0.6}$	$29.04_{\pm 0.4}$

Performance generally is lower on the transformed test sets

Experiments

Data Augmentation to create transformed test sets

Transformation	Example
Original	airforce 1 women shoes white
Shuffled	shoes women white airforce 1
Butterfinger	airvorce 1 women shoes white
Numeric	airforce 6 women shoes white
Color	airforce 1 women shoes green
Mention Replacement	zerogrand boys shoes leopard
All Transformations	shofs boys maple zerogrand

Training Data	Original			Transformed + Original Training		
Eval Data	Transformed Tests			Transformed Tests		
Augmentation	Precision	Recall	F1	Precision	Recall	F1
butterfinger	$30.86_{\pm 0.6}$	$33.65_{\pm 0.5}$	$32.19_{\pm 0.5}$	$54.59_{\pm 0.7}$	$54.88_{\pm 0.5}$	$54.73_{\pm 0.6}$
numeric	$60.80_{\pm 0.7}$	$58.32_{\pm 0.5}$	$59.54_{\pm 0.6}$	$60.45_{\pm 0.6}$	$59.32_{\pm 0.6}$	$59.88_{\pm 0.6}$
mention replacement	$74.83_{\pm 1.1}$	$75.52_{\pm 1.0}$	$75.17_{\pm 1.1}$	$74.88_{\pm 0.7}$	$77.74_{\pm 0.7}$	$76.28_{\pm 0.7}$
color	$61.55_{\pm 0.6}$	$60.62_{\pm 0.4}$	$61.08_{\pm 0.5}$	$61.37_{\pm 0.9}$	$60.59_{\pm 0.8}$	$60.98_{\pm 0.8}$
shuffled	$59.82_{\pm 0.5}$	$59.41_{\pm 0.5}$	$59.62_{\pm 0.4}$	$64.50_{\pm 0.4}$	$66.08_{\pm 0.5}$	$65.28_{\pm 0.4}$
all transformations	$27.70_{\pm 0.4}$	$30.53_{\pm 0.6}$	$29.04_{\pm 0.4}$	$66.25_{\pm 0.7}$	$69.36_{\pm 0.6}$	$67.77_{\pm 0.6}$

But improves if trained on augmented data

Brandeis

Experiments

And still performs well on the original test set

Training Data	Original			Transformed + Original Training			Transformed + Original Training		
Eval Data	Transformed Tests			Transformed Tests			Original Test		
Augmentation	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
butterfinger	$30.86_{\pm 0.6}$	$33.65_{\pm 0.5}$	$32.19_{\pm 0.5}$	$54.59_{\pm 0.7}$	$54.88_{\pm 0.5}$	$54.73_{\pm 0.6}$	$58.93_{\pm 0.5}$	$59.78_{\pm 0.5}$	$59.35_{\pm 0.5}$
numeric	$60.80_{\pm 0.7}$	$58.32_{\pm 0.5}$	$59.54_{\pm 0.6}$	$60.45_{\pm 0.6}$	$59.32_{\pm 0.6}$	$59.88_{\pm 0.6}$	$60.41_{\pm 0.7}$	$59.35_{\pm 0.7}$	$59.88_{\pm 0.7}$
mention replacement	$74.83_{\pm 1.1}$	$75.52_{\pm 1.0}$	$75.17_{\pm 1.1}$	$74.88_{\pm 0.7}$	$77.74_{\pm 0.7}$	$76.28_{\pm 0.7}$	$60.04_{\pm 0.8}$	$59.50_{\pm 0.7}$	$59.77_{\pm 0.7}$
color	$61.55_{\pm 0.6}$	$60.62_{\pm 0.4}$	$61.08_{\pm 0.5}$	$61.37_{\pm 0.9}$	$60.59_{\pm 0.8}$	$60.98_{\pm 0.8}$	$60.58_{\pm 0.7}$	$59.05_{\pm 0.6}$	$59.80_{\pm 0.6}$
shuffled	$59.82_{\pm 0.5}$	$59.41_{\pm 0.5}$	$59.62_{\pm 0.4}$	$64.50_{\pm 0.4}$	$66.08_{\pm 0.5}$	$65.28_{\pm 0.4}$	$60.32_{\pm 0.5}$	$59.97_{\pm 0.6}$	$60.14_{\pm 0.5}$
all transformations	$27.70_{\pm 0.4}$	$30.53_{\pm 0.6}$	$29.04_{\pm 0.4}$	$66.25_{\pm 0.7}$	$69.36_{\pm 0.6}$	$67.77_{\pm 0.6}$	$55.65_{\pm 0.7}$	$54.74_{\pm 0.7}$	$55.19_{\pm 0.7}$

Limitations & Future Work

- Ontology has broad coverage, but may have different distributions of types across other datasets
- Most of annotator disagreement is in spans
- English only currently
- Potential for future expansion to other languages
  - Amazon ESCI dataset has Spanish and Japanese

Conclusion

- QueryNER: E-commerce query segmentation dataset
- Ontology of 17 entity types
- Dataset of nearly 10,000 human annotated queries
- Showed simple data augmentation strategies can make segmentation model more robust

Available at: github.com/bltlab/query-ner