amazon

AMAZON RESEARCH DAYS

A DIVERSE **NETWORK**
A UNITED COMMUNITY

2021 AMAZON RESEARCH DAYS

KUSHAL KUMAR

Applied Scientist

India Machine Learning

Solving Price Per Unit Problem Around the World: Formulating Fact Extraction as Question Answering

Tarik Arici, **Kushal Kumar**, Hayreddin Çeker,
Anoop S V K K Saladi, Ismail Tutar

What is PPU?

Featured from our brands

AmazonFresh Colombia Ground Coffee, Medium Roast, 12 Ounce

Ground · 12 Ounce (Pack of 1)

1,573

$5.3 ($0.44/Ounce)

Save more with Subscribe & Save

prime **Today 1PM - 6PM**

Sponsored

Gevalia Colombian Medium Roast Ground Coffee (12 oz Bags, Pack of 6)

Colombian · Ground · 12 Ounce (Pack of 6)

580

$35.88 ($0.50/Ounce)

Save 5% more with Subscribe & Save

Sponsored

Gevalia Colombian Blend Medium Roast K-Cup Coffee Pods (72 Pods,12 Count Pack of 6)

Pods · 12 Count (Pack of 6)

337

$48.5 ($0.67/Count)

Save 5% more with Subscribe & Save

Featured from our brands

Amazon Brand - 100 Ct Solimo Dark Roast Coffee Pods, Compatible with Keurig 2.0 K-Cup Brewers

Dark Roast · Pods · 100 Count (Pack of 1)

26,987

$29.9 ($0.30/Count)

Save 5% more with Subscribe & Save

Challenges with PPU

- Sellers put wrong PPU info
- Specify it in free-text
- Information in product image

Typhoo Organic Herbal Infusion Slim Tea 20 Tea Bag Pack Of 2

Imported 10 Colour Extra Flame Beauty Eyeshadow (Assorted) Buy 1 Get 1 Free with Kajal

Herbal Strategi Mosquito Repellent Agarbatti (3 x 40 Sticks) (Pack of 3)

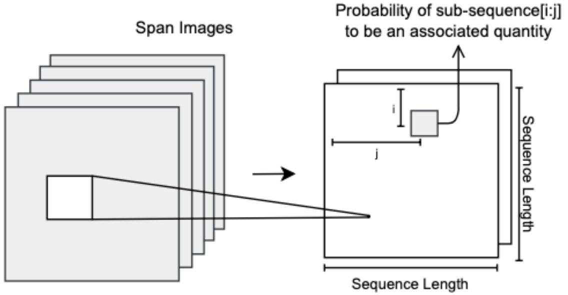
Challenges with PPU

- Sellers put wrong PPU info
- Specify it in free-text
- Information in product image
- Unstructured
- Extracting only relevant values
- Computing total quantity

Girnar Instant Tea Combo of Masala and Cardamon Premix (140g)

Problem Formulation

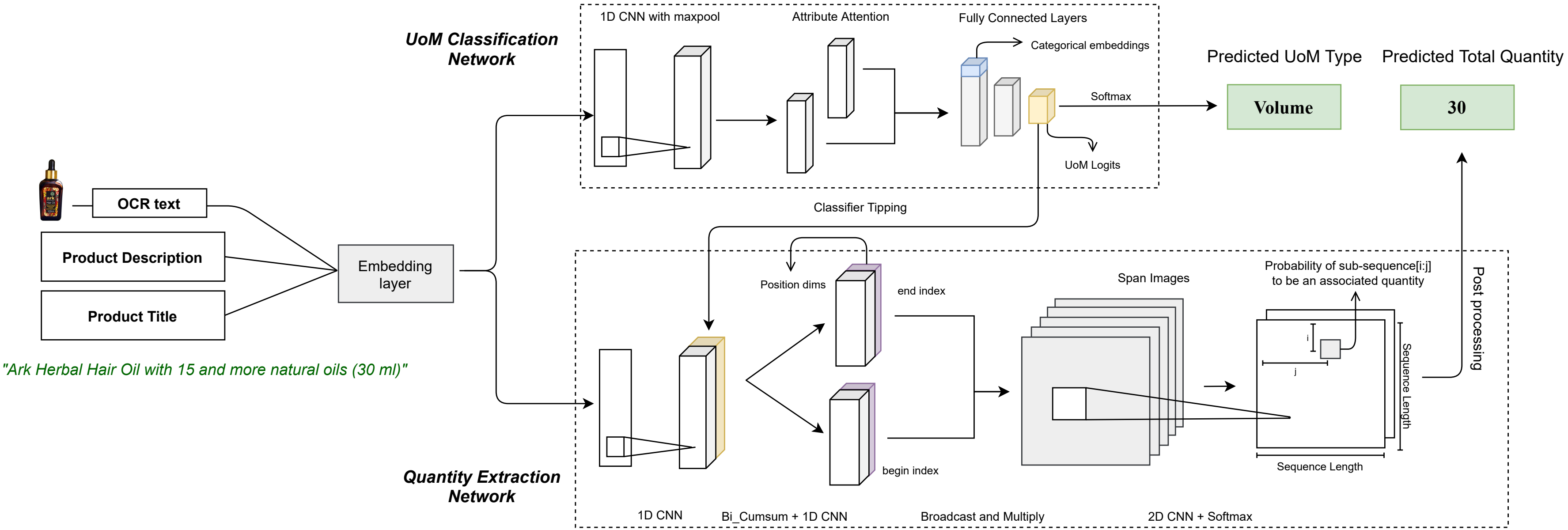
- Named Entity Recognition (NER) is a de-facto solution to fact extraction. However, it does not allow to couple start and end indices explicitly. "*fluid ounce*" or "*fl oz*" need to be tagged.
- A significant advance in answer span prediction is the BiDAF method which uses bi-directional LSTM on query-to-context and context-to-query sequences followed by softmax normalization across sequence dimension.
- To overcome these limitations, we introduce a span-image architecture that works at a character-level and employ a QA approach to quantity extraction which conditions the extractor model with UoM type specific question (*eg. "What is the total volume?"*).



Problem Formulation

- Named Entity Recognition (NER) is a de-facto solution to fact extraction. However, it does not allow to couple start and end indices explicitly. "*fluid ounce*" or "*fl oz*" need to be tagged.
- A significant advance in answer span prediction is the BiDAF method which uses bi-directional LSTM on query-to-context and context-to-query sequences followed by softmax normalization across sequence dimension.
- To overcome these limitations, we introduce a span-image architecture that works at a character-level and employ a QA approach to quantity extraction which conditions the extractor model with UoM type specific question (*eg.* "*What is the total volume?*").
- Formally, for a product P with set of N text attributes - $\{a_1^P, a_1^P, \dots a_N^P\}$, with character sequences $x^i = \{x_1^i, x_2^i, \dots x_{n_i}^i\}$, where $x_j^i \in \mathbb{R}^k$ are k-dimensional character embedding for j^{th} character in the sequence of length n_i for attribute a_i^P, task is to predict UoM type and use it to predict the begin and end indices of all relevant quantities in each x^i.

Our Model Architecture



Dataset

- Use auditors for product labels
- Tag tokens using some heuristics
- Add noise
  - adding, deleting, changing characters
  - adding gibberish words
  - adding token relevant to UoM task

Annotate product with UoM Type and Quantity

Obtain token-level tags for training

Improve model generalizability

TABLE II: Distribution of products by number of spans

# of spans	Percentage
0	54.0%
1	34.5%
2	11.3%
3	0.2%

Comparison with baselines

Comparison with Rule-based models

TABLE III: Performance gains for our PPU model over the rule-based model in US store

Task		PPU model (Δ)			
		Volume	Weight	Count	Overall
UoM Classification	ΔP	1.6	26.0	17.9	16.4
	ΔR	64.0	56.9	90.0	76.3
	ΔF1	**51.6**	**44.1**	**86.8**	**65.4**
Quantity Extraction	ΔP	**11.5**	**37.4**	**42.2**	**34.4**
	ΔR	19.1	0.2	27.7	19.1
	ΔF1	22.4	8.0	40.6	26.1

- Our model outperforms rule based models for both the tasks across UoM types
- Rule-based models fail to capture complex patterns due to lack of semantic understanding of the product
- Rule-based model lacks scalability unlike DL based model which can improve with more training data.

Comparison with BERT-based models

Our model outperforms BERT-based sequence classifier model in US store

It comes really close in precision on IN store, while BERT outperforms our model in EU-5

Despite using a light-weight model, our CNN architecture is deep enough to learn semantic information required for the UoM task.

TABLE IV: Performance gains over pre-trained Google BERT model fine-tuned on catalog for both MLM and NSP tasks and on UoM Classification task across all stores

Store		PPU Model Classifier (Δ)			
		Volume	Weight	Count	Overall
EU-5	ΔP	-11.8	-20.8	-7.2	-12.0
	ΔR	-9.0	-7.8	-3.9	-5.4
	ΔF1	-10.2	-13.5	-5.6	-8.6
IN	ΔP	0.0	0.0	0.0	0.0
	ΔR	-8.0	-4.4	-3.9	-5.0
	ΔF1	-4.4	-2.3	-2.1	-2.7
US	ΔP	0.5	-5.0	0.3	-0.7
	ΔR	5.8	12.3	3.6	6.8
	ΔF1	**3.3**	**3.7**	**2.0**	**3.2**

Comparison with BERT-based models

QA outperforms FE approach even with bulkier language model in US store with 10.6% lift in precision and EU store with 0.9% lift in precision

We compare precision as it is a key for deployment, our model crosses the bar in US and also 2 out of 3 UoM types in EU-5 and IN

However, fact extraction approach crosses the set precision bar only for 2 out of 3 UoM types in US and IN store, and none in EU-5

TABLE V: Problem formulation - Performance gains for Question Answering approach using our PPU model over Fact Extraction approach using BERT across all stores

Store		Question Prediction and Answering (Δ)			
		Volume	Weight	Count	Overall
EU-5	ΔP	**1.7**	**1.2**	**0.3**	**0.9**
	ΔR	-27.3	-20.2	-15.3	-19.6
	ΔF1	-30.7	-25.7	-17.2	-22.4
IN	ΔP	**0.1**	**0.1**	-0.6	**0.1**
	ΔR	-14.1	-7.4	11.3	-0.5
	ΔF1	-8.7	-4.7	10.8	1.8
US	ΔP	-0.3	-0.1	**19.8**	**10.6**
	ΔR	-8.3	-4.8	8.5	1.6
	ΔF1	-7.6	-5.0	12.0	3.8

Comparison On Latency Scaling

TABLE VII: Latency scaling (in milliseconds) with respect to number of CPU cores for different models.

Number of CPU cores	BERT		PPU model		PPU model - IN	
	Mean	P90	Mean	P90	Mean	P90
2	126	150	104	205	12	21
4	73	89	56	105	8	13
8	69	86	33	56	7	9
16	56	66	21	32	6	7

- Our model scales well on latency, where it achieves <50 ms latency on a machine with 8 cores
- BERT based models fail to achieve <50 ms latency even on a machine twice the size
- Removing noisy long-text attributes in certain stores further help improve model latency

Conclusion

1. We presented a lightweight deep learning model that can
   a) perform semantic learning for multi- answering,
   b) scale with our dataset sizes and
   c) be shared across stores, thanks to its fully character based architecture

2. UoM for computing quantity depends on factors such as brand, product type, conventions in a store etc. As human labelled data is limited in size, large language models are not the best fit.

3. This fully character based architectures desirable, while span-image allows multi-answering.

4. Solving quantity extraction as a question prediction and answering task gives better performance over fact extraction formulation even with bulkier pre-trained language models like BERT.

2021 AMAZON RESEARCH DAYS

Thank You!