CRAC 2022

Analyzing Coreference and Bridging in Product Reviews

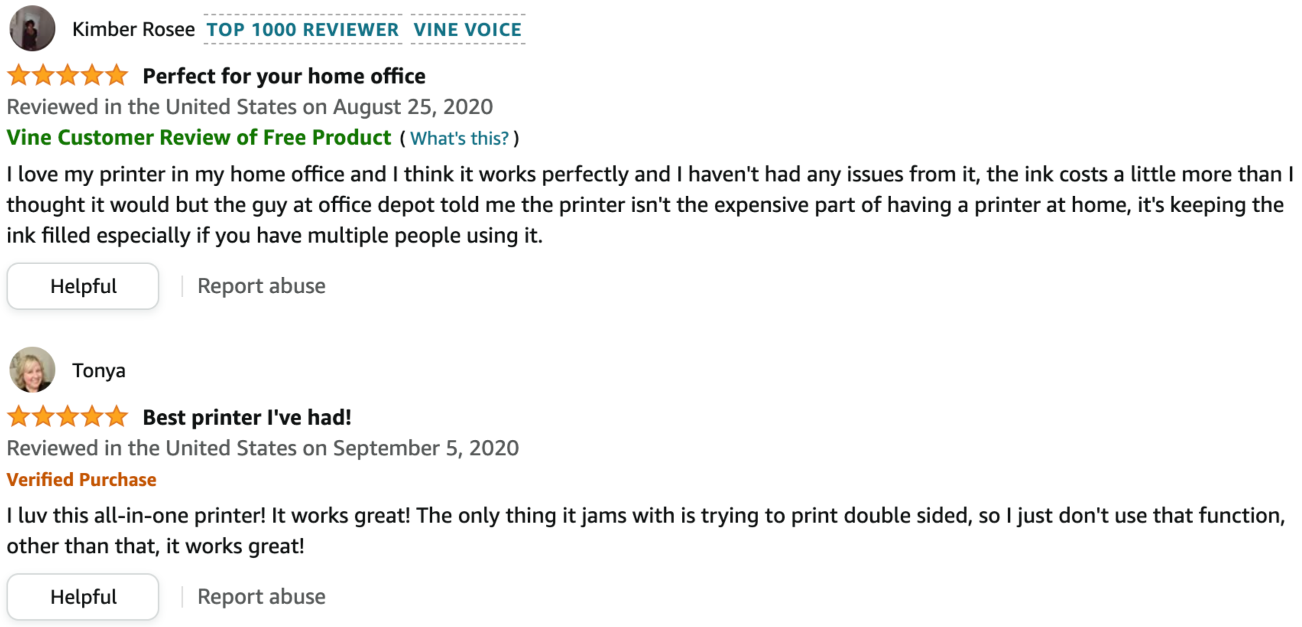
Hideo Kobayashi, University of Texas at Dallas

Christopher Malon, NEC Laboratories America

Towards Automated Understanding of Product Reviews

- Aspect based sentiment analysis (ABSA), opinion summarization in product reviews
  - factuality checking to make sure summaries are correct

- Existing ABSA and factuality scores do not learn to catch coreference or bridging errors



Kimber Rosee TOP 1000 REVIEWER VINE VOICE

★★★★★ **Perfect for your home office**

Reviewed in the United States on August 25, 2020

Vine Customer Review of Free Product (What's this?)

I love my printer in my home office and I think it works perfectly and I haven't had any issues from it, the ink costs a little more than I thought it would but the guy at office depot told me the printer isn't the expensive part of having a printer at home, it's keeping the ink filled especially if you have multiple people using it.

Tonya

★★★★★ **Best printer I've had!**

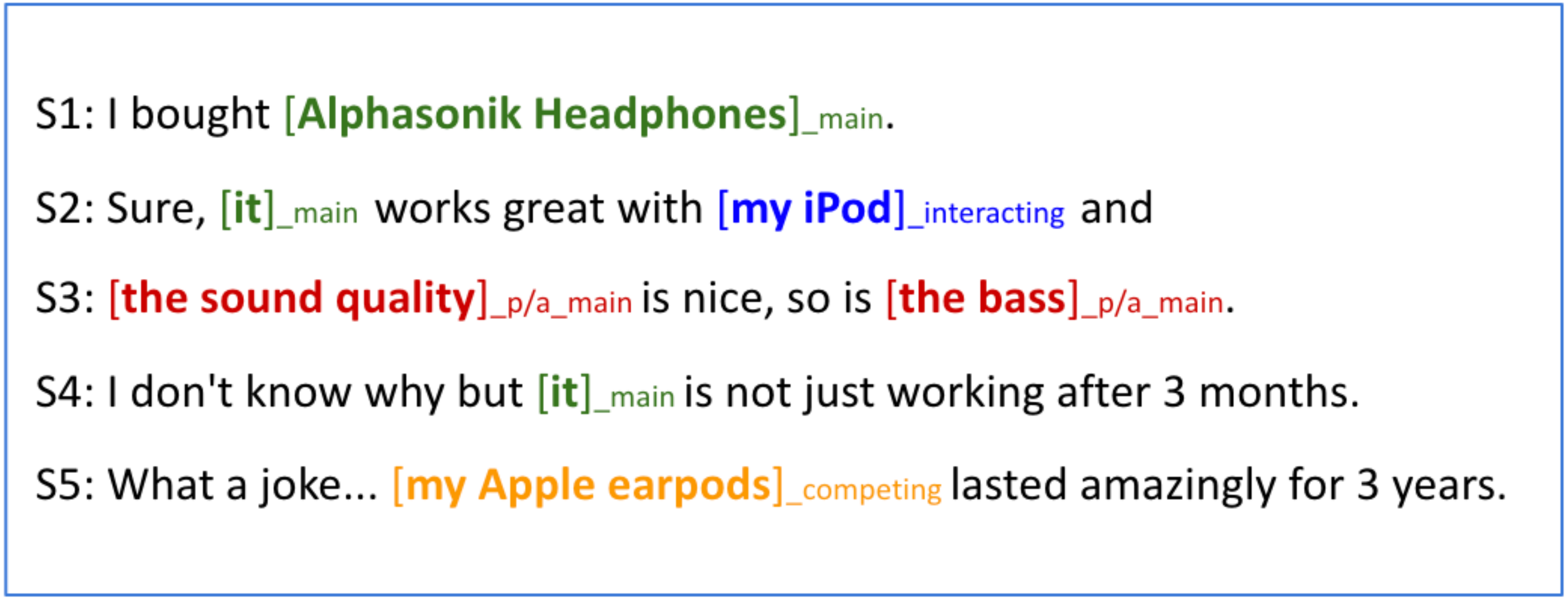
Reviewed in the United States on September 5, 2020

Verified Purchase

I luv this all-in-one printer! It works great! The only thing it jams with is trying to print double sided, so I just don't use that function, other than that, it works great!

Technical Challenge: Understanding complex reviews

- Need to understand entity ambiguities
  - Identical to what? (Coreference)
  - Part or attribute of what?: (Bridging)



S1: I bought [Alphasonik Headphones]_main.

S2: Sure, [it]_main works great with [my iPod]_interacting and

S3: [the sound quality]_p/a_main is nice, so is [the bass]_p/a_main.

S4: I don't know why but [it]_main is not just working after 3 months.

S5: What a joke... [my Apple earpods]_competing lasted amazingly for 3 years.

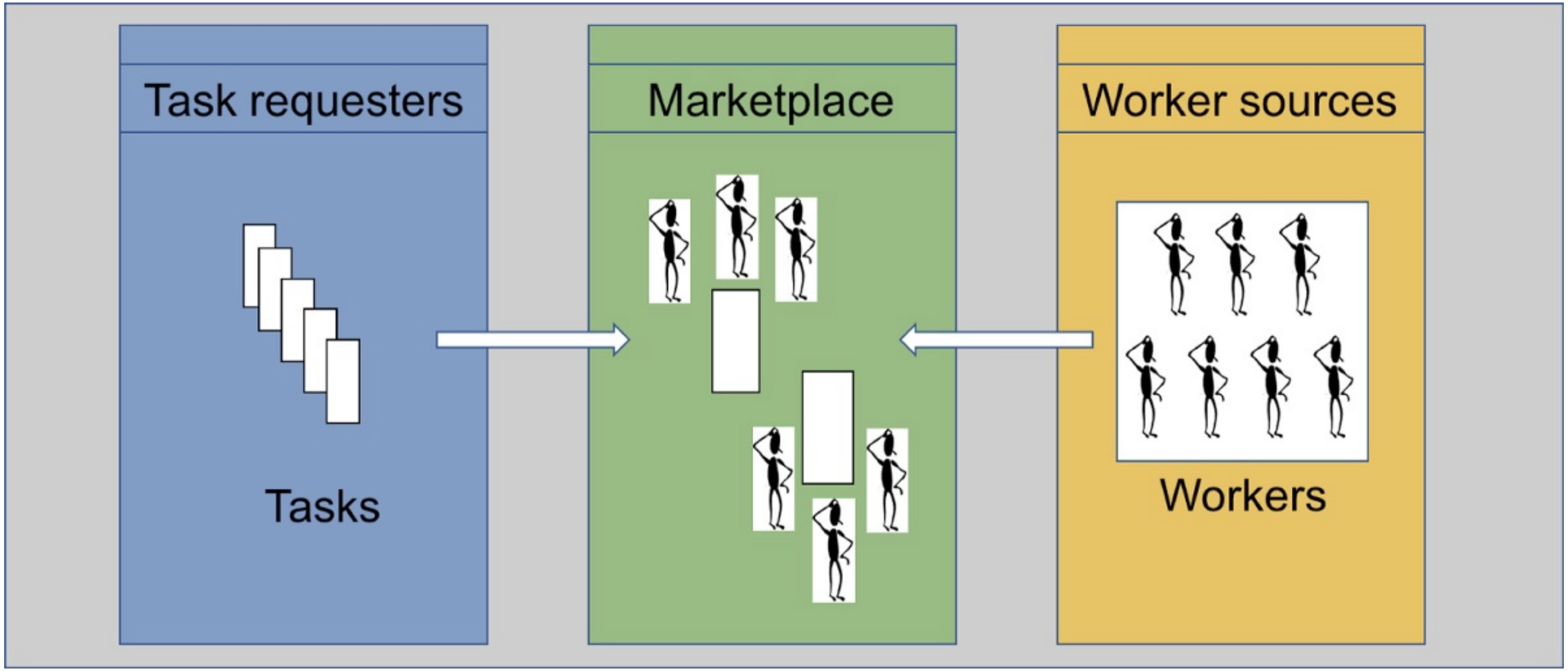
But... annotating coreference and bridging is expensive

Contributions

- Define Mention classification task: Annotation scheme & Crowdsourced dataset creation
- Analyze an existing coreference system in product reviews
- Create NLI test set & Show the weakness of a SOTA factuality checking (NLI)

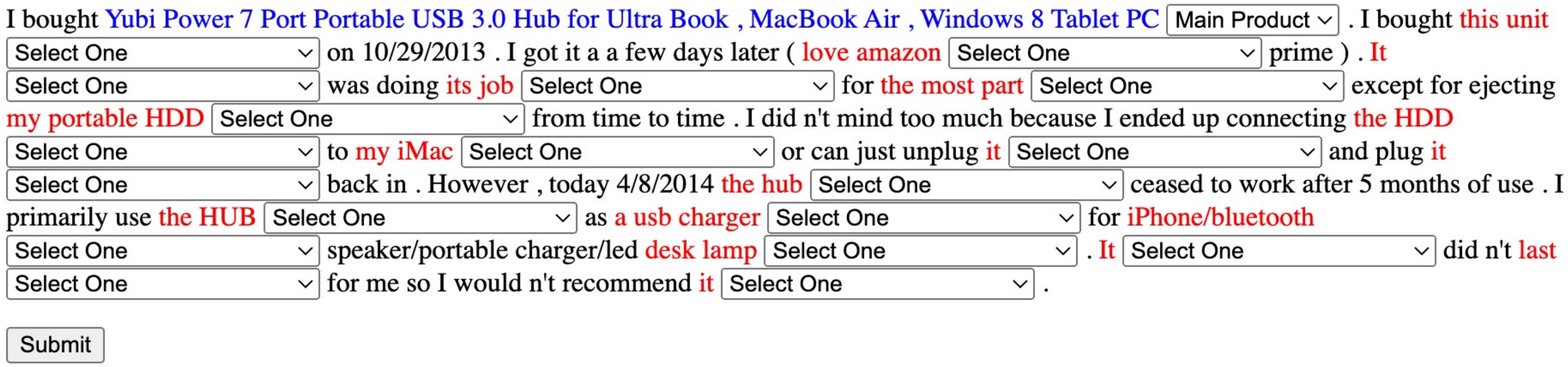
Crowdsourced Dataset Creation

- Annotated 498 Amazon electronics reviews via Amazon Mechanical Turk



Crowdsourced Dataset Creation

- Annotated 498 Amazon electronics reviews via Amazon Mechanical Turk
- Annotation: Identical? Part/Attribute of?
  - Main product
  - Competing product
  - Product interacting with the main product or competitors
  - Generic term for the category of the main product



I bought Yubi Power 7 Port Portable USB 3.0 Hub for Ultra Book , MacBook Air , Windows 8 Tablet PC Main Product . I bought this unit Select One on 10/29/2013 . I got it a a few days later (love amazon Select One prime) . It Select One was doing its job Select One for the most part Select One except for ejecting my portable HDD Select One from time to time . I did n't mind too much because I ended up connecting the HDD Select One to my iMac Select One or can just unplug it Select One and plug it Select One back in . However , today 4/8/2014 the hub Select One ceased to work after 5 months of use . I primarily use the HUB Select One as a usb charger Select One for iPhone/bluetooth Select One speaker/portable charger/led desk lamp Select One . It Select One did n't last Select One for me so I would n't recommend it Select One .

Resulting Dataset

- We use Cohen’s kappa (Cohen, 1960) to measure inter-annotator agreement
- Agreement is **substantial**: kappa is 0.681

Mention Type	Counts
Main	2864
P/A of Main	1512
Competing	429
P/A of Competing	103
Generic	193
P/A of Generic	18
Interacting	853
P/A of Interacting	308
Others	2127

Annotated 8,894 mentions with substantial agreements

Resulting Dataset: Confusion Matrix for Annotation Agreements

- Many generic mentions are thought to refer to the main product
- Part/attribute of a generic mention may be confused with a main or competing product

	Main	P/A of Main	Com	P/A of Com	Gen	P/A of Gen	Int	P/A of Int	Oth
Main	94.8	1.9	1.3	0.1	**7.4**	0.3	0.9	0.3	1.3
P/A of Main	2.0	88.0	0.5	5.0	1.7	**10.2**	1.4	6.0	4.9
Com	0.4	0.1	89.1	2.0	5.1	1.3	0.6	0.4	0.5
P/A of Com	0.0	0.4	1.2	80.3	0.3	**7.6**	0.3	0.5	0.5
Gen	0.8	0.3	3.3	0.6	81.3	2.3	0.5	0.3	0.3
P/A of Gen	0.0	0.4	0.1	3.3	0.9	73.3	0.1	0.2	0.2
Int	0.4	1.3	0.9	0.6	0.9	1.7	89.1	7.7	2.0
P/A of Int	0.1	1.5	0.6	2.0	0.3	1.0	3.2	79.0	1.4
Oth	1.4	6.2	3.0	6.0	2.2	2.3	4.0	5.7	88.9

Overall, annotation scheme was clear to workers

Results: Evaluating SOTA Coreference Resolver in Product Reviews

- Pretrained Coreference (Xu and Choi, 2020) underperforms in out-of-domain (i.e., product domains)

	MUC			B3			CEAFφ4			
	P	R	F1	P	R	F1	P	R	F1	AVG F1
OntoNotes	85.9	85.5	85.7	79	78.9	79	76.7	75.2	75.9	80.2
Main	68.3	59.5	63.6	63.1	48.3	54.7	50.5	68.1	58.0	58.8
Competing	37.1	27.4	31.6	43.7	28.8	34.7	57.7	40.6	47.7	38.0
Generic	22.2	11.8	15.4	32.3	14.0	55.0	19.6	18.8	28.0	21.0

Weakness of SOTA Factuality Checking: System

- Task: Given a hypothesis and a source review, classify if a hypothesis is consistent with the review.

- Question: Do factuality scores detect coreference errors?

- SummaC-ZS (Laban et al., 2022)
  - NLI-based
  - Zero-shot

Weakness of SOTA Factuality Checking: Test Set Creation

- For the mention categories “Main product,” “Competing product,” and “Interacting product,” take sentences that contain the second or subsequent mentions of these categories

- Construct one sentence in which we replace that mention with the main product name, or the first mention of a competing product, or the first mention of an interacting product.

- Manually check labels

Weakness of SOTA Factuality Checking: Test Set Creation

- Replacing competing product by main product

  Review: "… My associate and I played with it for a couple days trying to get the video to be in focus but we never got it to look right. I bought a Flip and **it** worked great. Sadly the Flip used AA batteries and was more expensive but at least the video was in focus…"

  Hypothesis: I bought a Flip and **Creative Labs Vado Pocket Video Camcorder** worked great.

  Human judgment: Inconsistent

Results: Weakness of Factuality Checking: Results

- Inconsistent substitutions are mostly not caught

Original	Replacement	Label	Accuracy
Main	Main	Consis.	100%
Main	Competing	Inconsis.	**20%**
Main	Interacting	Inconsis.	**38%**
Competing	Competing	Consis.	87%
Competing	Main	Inconsis.	**44%**
Competing	Interacting	Inconsis.	**50%**
Interacting	Interacting	Consis.	89%
Interacting	Main	Inconsis.	**32%**
Interacting	Competing	Inconsis.	100%

Significant room for improvement in distinguishing non-identical entities

Future Work

- We completed
  - Defining Mention classification task: Annotation scheme & Crowdsourced dataset creation
  - Analyzing SOTA coreference system in product reviews
  - Creating NLI test set & Showing the weakness of a SOTA consistency detection

- Next steps are
  - Collecting more data via crowdsourcing platform
  - Training a mention type classifier
  - Analyzing SOTA bridging system in product reviews
  - Integrating mention information into the factuality checking NLI system

Q&A