Recommend for a Reason: Unlocking the Power of Unsupervised Aspect-Sentiment Co-Extraction

Zeyu Li[1], Wei Cheng[2], Reema Kshetramade[1], John Houser[1], Haifeng Chen[2], Wei Wang[1]

[1]Department of Computer Science, University of California, Los Angeles, [2]NEC Labs America

EMNLP 2021 Online & in the Dominican Republic
Conference on Empirical Methods in Natural Language Processing

1. Introduction

- *Compliments* and *concerns* in reviews are valuable for understanding users' shopping interests and their **opinions** to specific **aspects** of items.
- *Existing work*: ignores the fact that users may hold different **attentions** to various **properties** of the merchandise – they may show strong attentions to certain properties but indifference to others. (see below example)

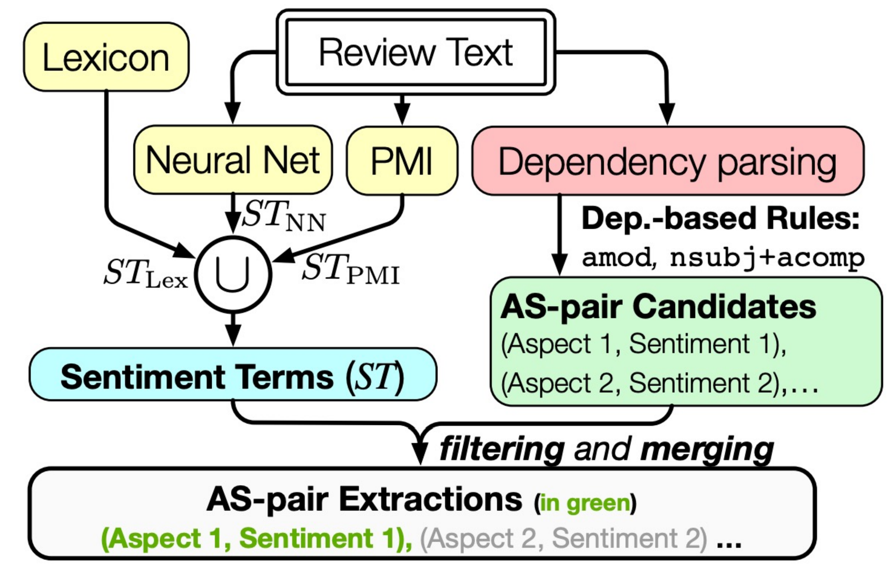
Reviews	Microphone	Comfort	Sound
R1 [5 stars]: *Comfortable. Very high quality sound. …Mic is good too. There is an switch to mute your mic… I wear glasses and these are comfortable with my glasses on. …*	*good* (satisfied)	*comfortable*	*high quality* (praising)
R2 [3 stars]: *I love the comfort, sound, and style but the mic is complete junk!*	*complete junk* (angry)	*love*	*love*
R3 [5 stars]: *…But this one feels like a pillow, there's nothing wrong with the audio and it does the job. …con is that the included microphone is pretty bad.*	*pretty bad* (unsatisfied)	*like a pillow* (enjoyable)	*nothing wrong*

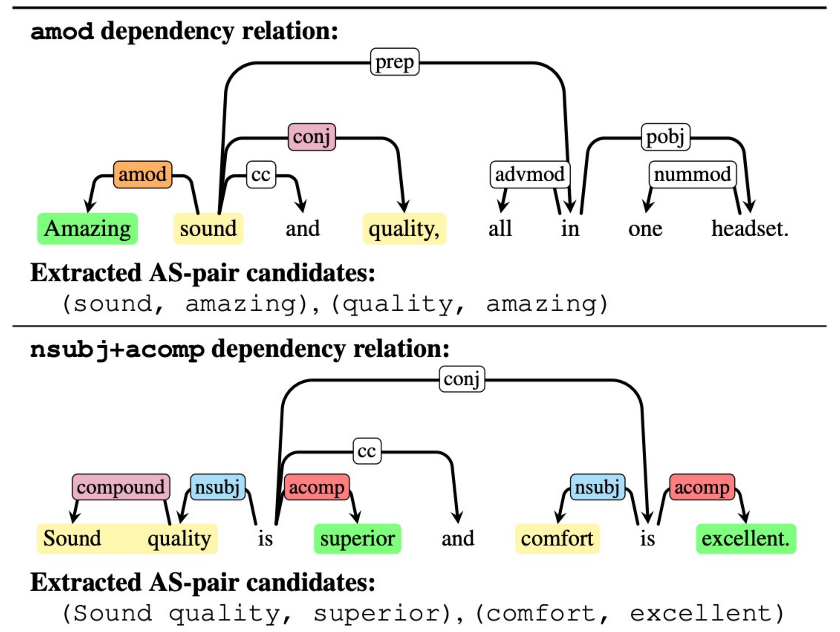
Table 1: Example reviews of a headset with three aspects, namely microphone quality, comfort level, and sound quality, highlighted specifically. The extracted sentiments are on the right. R1 vs. R2: Different users react differently (microphone quality) to the same item due to distinct personal attentions and, consequently, give divergent ratings. R1 vs. R3: A user can still rate highly of an item due to special attention on particular aspects (comfort level) regardless of certain unsatisfactory or indifferent properties (microphone and sound qualities).

- Use NLP methods to extract explicit and definitive ***sentiment and aspect terms*** which models user attentions and item properties.
- We propose a tightly coupled two-stage approach: *Aspect-Sentiment Pair Extractor (**ASPE**) + Attention-Property-aware Rating Estimator (**APRE**).*

2. ASPE

- [First Step] Sentiment terms extraction:
  - Pointwise Mutual Information-based (PMI-based) (Polarity)
  - Neural Network-based (NN-based) (Linguistic patterns)
  - Knowledge/Lexicon-based (Existing knowledge)
  - $ST = ST_{PMI} \cup ST_{NN} \cup ST_{Lex}$





- [Second Step] AS-pair extraction
  - First: label AS-pair candidates using dependency parsing
    - amod and nsubj+acomp
  - Second: filter out non-sentiment-carrying candidates using ST

3. APRE

- Language encoding with pre-trained BERT
- **Explicit** aspect-level attitude modeling

$$\alpha_{u,r}^{(a)} = \frac{\exp(\tanh(\boldsymbol{w}_{\text{ex}}^T[\boldsymbol{h}_{u,r}^{(a)};\boldsymbol{a}^{(u)}]))}{\sum_{r'\in R^u}\exp(\tanh(\boldsymbol{w}_{\text{ex}}^T[\boldsymbol{h}_{u,r'}^{(a)};\boldsymbol{a}^{(u)}]))} \quad \boldsymbol{g}_u^{(a)} = \sum_{r\in R^u}\alpha_{u,r}^{(a)}\boldsymbol{h}_{u,r}^{(a)}$$

- **Implicit** review representation

$$\boldsymbol{v}_{u,r} = \left[\boldsymbol{h}_{\text{[CLS]}}^1;\boldsymbol{h}_{\text{cnn}};\text{MaxPool}(\mathbf{H}^1);\text{AvgPool}(\mathbf{H}^1)\right]$$

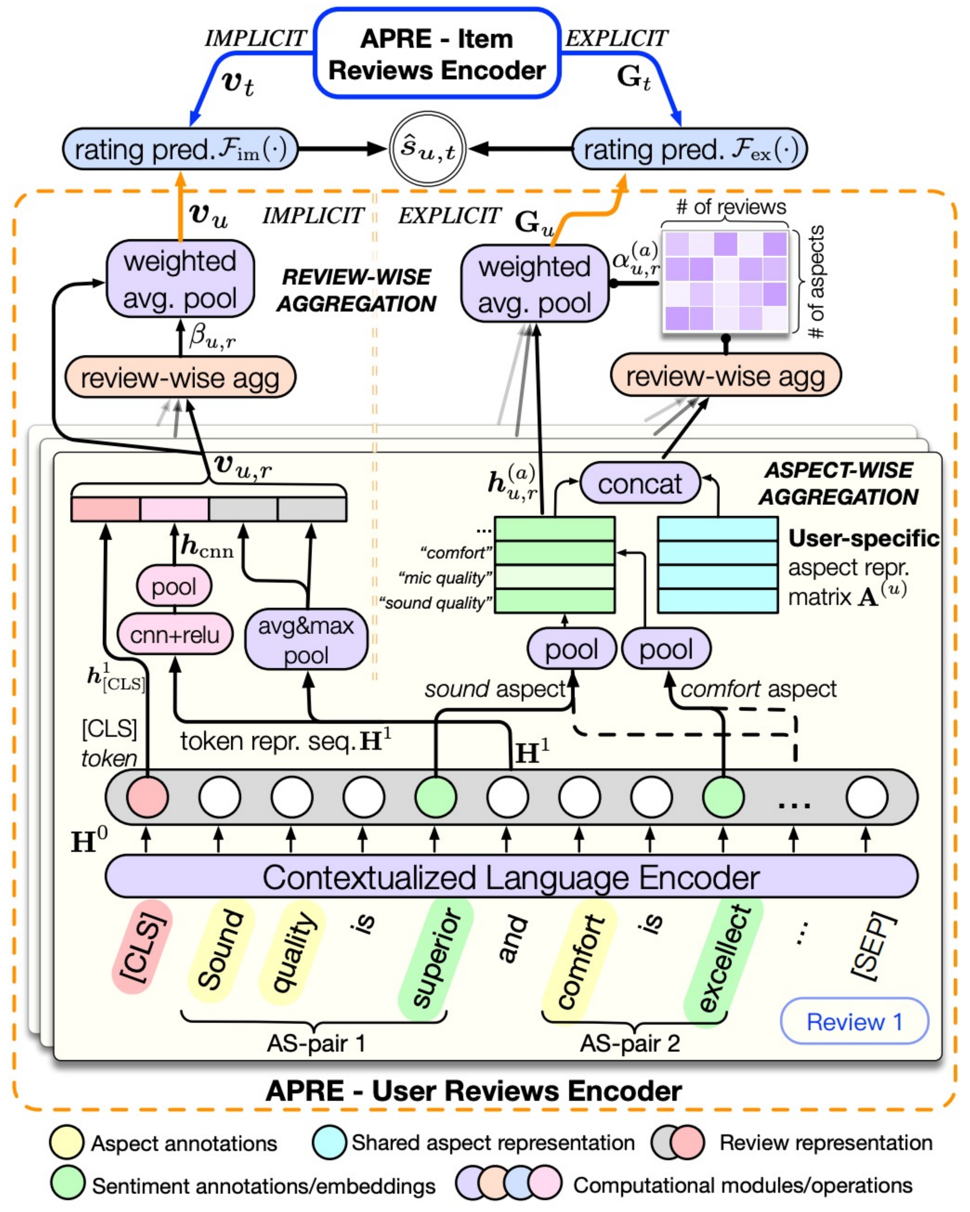
$$\boldsymbol{h}_{\text{cnn}} = \text{MaxPool}(\text{ReLU}(\text{ConvNN\_1D}(\mathbf{H}^1))).$$

$$\beta_{u,r} = \frac{\exp(\tanh(\boldsymbol{w}_{\text{im}}^T\boldsymbol{v}_{u,r}))}{\sum_{r'\in R^u}\exp(\tanh(\boldsymbol{w}_{\text{im}}^T\boldsymbol{v}_{u,r'}))} \quad \boldsymbol{v}_u = \sum_{r\in R^u}\beta_{u,r}\boldsymbol{v}_{u,r}$$

- Rating regression and optimization

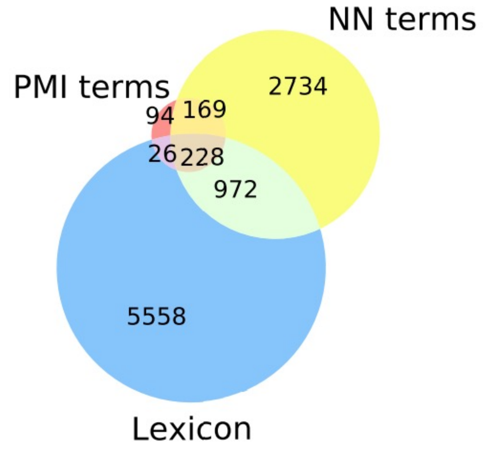
$$\hat{s}_{u,t} = \underbrace{b_u + b_t}_{\text{biases}} + \underbrace{\mathcal{F}_{\text{im}}([\boldsymbol{v}_u;\boldsymbol{v}_t])}_{\text{implicit feature}} + \underbrace{\langle\boldsymbol{\gamma},\mathcal{F}_{\text{ex}}([\mathbf{G}_u;\mathbf{G}_t])\rangle}_{\text{explicit feature}}$$

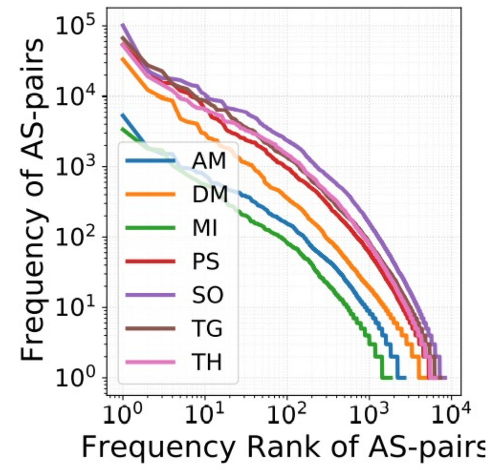
$$J(\Theta) = \sum_{r_{u,t}\in R}(s_{u,t}-\hat{s}_{u,t})^2 + L_2\text{-reg}(\lambda)$$



4. Experiments

- Datasets: 7 Amazon Review Datasets. (8:1:1)
- Baseline models: 13 models (traditional or deep learning category)
- Evaluation metric: MSE on test set
- **Results:** AS-pair Extraction of ASPE





AM	DM	MI	PS	SO	TG	TH
ItemTok	song	*ItemTok*	*ItemTok*	*ItemTok*	*ItemTok*	*ItemTok*
product	*ItemTok*	sound	dog	knife	toy	light
time	album	guitar	food	quality	game	tool
car	music	string	cat	product	piece	quality
look	time	quality	toy	size	quality	price
price	sound	tone	time	price	child	product
quality	voice	price	product	look	color	bulb
light	track	pedal	price	bag	part	battery
oil	lyric	tuner	treat	fit	fun	size
battery	version	cable	water	light	size	flashlight

- The distributions of the frequencies AS-pairs follow the Zipf's Law meaning that ASPE performs **consistently** across domains.

- **Results:** Rating Prediction of APRE
- Demonstrates the superior capability of APRE to make accurate rating predictions in different domains (Ours vs. the rest); The performance improvement is NOT because of the use of BERT (Ours vs. AHN-B)
- **Ablation Study** reflects positive contributions from both channels

Models	AM	DM	MI	PS	SO	TG	TH
TRADITIONAL MODELS							
MF	1.986	1.715	2.085	2.048	2.084	1.471	1.631
WRMF	1.327	0.537	1.358	1.629	1.371	1.068	1.216
FM	1.082	0.436	1.146	1.458	1.212	0.922	1.050
DEEP LEARNING-BASED MODELS							
ConvMF	1.046	0.407	1.075	1.458	1.026	0.986	1.104
NeuMF	0.901	0.396	0.903	1.294	0.893	0.841	1.072
D-Attn	0.816	0.403	0.835	1.264	0.897	0.887	0.980
D-CNN	0.809	0.390	0.861	1.250	0.894	0.835	0.975
NARRE	0.826	0.374	0.837	1.425	0.990	0.908	0.958
MPCN	0.815	0.447	0.842	1.300	0.929	0.898	0.969
ANR	0.806	0.381	0.845	1.327	0.906	0.844	0.981
DAML	0.829	0.372	0.837	1.247	0.893	0.820	0.962
AHN-B	0.810	0.385	0.840	1.270	0.896	0.829	0.976
AHN	0.802	0.376	0.834	1.252	0.887	0.822	0.967
OUR MODELS AND PERCENTAGE IMPROVEMENTS							
Ours	**0.791**	**0.359**	**0.823**	**1.218**	**0.863**	**0.788**	**0.936**
Δ(%)	1.390	3.621	1.337	2.381	2.784	4.061	2.350
Val.	0.790	0.362	0.821	1.216	0.860	0.790	0.933
ABLATION STUDIES							
w/o EX	0.814	0.379	0.833	1.244.	0.882	0.796	0.965
w/o IM	0.798	0.374	0.863	1.226	0.873	0.798	0.956

Aspects	material	smell	battery	install	look	price	sound
Attn. of u_*	✓	✓	✓	✓	✓	n/a	n/a
Prop. of t_*	n/a	n/a	✗	✓	✓	✗	✓/✗
Inferred Impact	*Unk.*	*Unk.*	***Neg.***	***Pos.***	***Pos.***	*Unk.*	*Unk.*
$\gamma_i\mathcal{F}_{\text{ex}}(\cdot)_i$ ($\times 10^{-2}$)	1.0	0.8	-0.8	1.9	1.5	0.9	0.6

From reviews given by user u_*. All aspects attended (✓).	
battery	**[To t_1]** After leaving this attached to my car for two days of non-use I have a ***dead*** battery. Never had a ***dead*** battery …, so *I am blaming this device.*
install	**[To t_2]** *This* was ***unbelievably easy*** to install. I have done …. The real key …the *installation* is so ***easy***. **[To t_3]** There were ***many*** *installation options*, but once …, *they clicked on easily.*
look	**[To t_3]** It was *not perfect and not shiny*, but it did *look* ***better***. **[To t_4]** It takes some ***elbow*** *grease*, but the *results are remarkable.*
material	**[To t_5]** *The plastic* however is ***very thin*** and *the cap* is ***pretty cheap***. **[To t_6]** Great value. …. They are ***very hard*** *plastic*, so they don't mark up panels.
smell	**[To t_7]** This has a ***terrible*** *smell* that really lingers awhile. It goes on green. …
From reviews received by item t_*.	
battery	**[From u_1]** The reason this won't work on an iPhone 4 or …because it uses ***low*** *power* Bluetooth, …. (✗)
install	**[From u_2]** Your mileage and gas mileage and cost of fuel is tabulated for each trip- *Installation* is ***pretty simple*** - but it …. (✓)
look	**[From u_3]** Driving habits, fuel efficiency, and engine health are ***nice*** *features*. The *overall design* is ***nice*** and *easy to navigate*. (✓)
price	**[From u_4]** In fact, there are similar products to this available at a ***much lower*** *price* that do work with …(✗)
sound	**[From u_5]** The Link device makes an ***audible*** *sound* when you go over 70 mpg, brake hard, or accelerate too fast. (✓) **[From u_6]** Also, the *beep* the link device makes …sounds ***really cheapy***. (✗)

- **Case Study:**
  - Seven example aspects with all reviews mentioning those categories;
  - APRE measures the aspect-level contributions of user-attention and item-property interactions via this term: <γ, F_{ex}([G_u;G_t])>;
  - ***Inferred Impact*** row states the interactional effects of user attentions and item properties based on our assumption that *attended aspects bear stronger impacts to the final prediction*.
  - This process of decomposition is a great way to ***interpret*** model prediction on an aspect-level granularity