Extracting Aspect Specific Sentiment Expressions implying Negative Opinions

Arjun Mukherjee
Department of Computer Science, University of Houston
arjun@uh.edu

Abstract. Subjective expression extraction is a central problem in fine-grained sentiment analysis. Most existing works focus on generic subjective expression extraction as opposed to aspect specific opinion phrase extraction. Given the ever-growing product reviews domain, extracting aspect specific opinion phrases is important as it yields the key product issues that are often mentioned via phrases (e.g., "signal fades very quickly," "had to flash the firmware often"). In this paper, we solve the problem using a combination of generative and discriminative modeling. The generative model performs a first level processing facilitating (1) discovery of potential head aspects containing issues, (2) generation of a labeled dataset of issue phrases, and (3) feed latent semantic features to subsequent discriminative modeling. We then employ discriminative large-margin and sequence modeling with pivot features for issue sentence classification and issue phrase boundary extraction. Experimental results using real-world reviews from Amazon.com demonstrate the effectiveness of the proposed approach.

1 Introduction

Aspect-based sentiment analysis is one of the main frameworks in opinion mining [1]. This thread focuses on unigram modeling as opposed to phrases which are more expressive. The fine-grained sentiment analysis paradigm [2] focuses on generic expressions as opposed to aspect specific expressions. Thus, there lies a big disconnect: *extracting aspect specific sentiment expressions* (*opinion phrases*).

Working in the most ubiquitous consumer reviews domain, this paper proposes a framework for extracting aspect specific opinion phrases. Further, we focus on sentiment expressions implying negative opinions. We call these *issues*. Extracting phrasal issues is important as they delineate the key problematic aspects of products that people want to know before making purchase decisions. Also, in contrast to positive opinion phrases that are relatively easier to discover (as they often involve direct positive opinions), discovering phrasal issues is more challenging as they appear in myriad types: direct ("signal strength was bad") or indirect ("had to flash firmware everyday"), containing verb phrase ("has been dropping connection"), noun, adjective or adverbial phrase ("voice commands operate only a limited set of features"), etc. We propose a holistic approach that caters for all types. Our approach is also context and polarity independent facilitating generic aspect specific opinion phrase extraction.

Formally, the task can be stated as follows: Given a sentence, $s = (w_1, ... w_n)$, discover the head aspect (HA/issue subject), $w_{HA=i}$ and a sub-sequence $(w_p ... w_q), p \leq i \leq s$ that best describes the issue, i.e., an aspect specific opinion phrase on the head

aspect and containing the head aspect. Throughout the paper, we will refer to head aspect and issue subject interchangeably. Examples below show labeled product issues within [[ ]] with the issue subject (head aspect) italicized:

- The first one I got working for about 2 weeks, then it [[started to drop the *signal*]], causing me to have to power cycle the unit.
- On the not so good side - We find the GPS [[*voice* to be not as clear]] as other GPSs we have used.

Although there are works that discover aspect/topic phrases using topic modeling [3, 4] and those that extract generic subjective expressions [5, 6, 7] using conditional random fields (CRFs), they lack the correspondence of the aspect and sentiment terms appearing in the sentence context. The proposed aspect specific sentiment expression task setting includes the head aspect within the phrase that naturally addresses the correspondence issue. Belonging to the family of information extraction problems, our task has resemblances with various works which are noted below.

In [8], subjective verb expressions were discovered using markov networks; in [9], supervised keyphrase extraction was used; and in [10], a re-ranking approach was used on the output of a sequence model to improve opinion expression extraction. These works mostly relied on word level features under the first-order Markov assumption. In [11, 12], segment features were used via semi-CRFs.

Parsing, phrasal, relational, and syntactic feature based approaches [13, 14, 15, 16] have also been successful in opinion mining. In our context, the work of [17] is relevant where features indicating dependency relations between opinion expressions were employed for opinion expression extraction. However, their approach relies on the output of a sequence labeler, prohibiting dependency features to be encoded in a sequence model. Other related works where sequence modeling was used include polarity identification [18, 19] and opinion relation extraction [20].

On a broader scope, this work is also related to the family of approaches in paraphrase learning [21], clustering [22], emotional paraphrase extraction [23], and keyphrase extraction [24] as they also discover phrase boundaries in their relevant contexts.

However, above works focus either on generic subjective expressions, aspect/topic phrases, paraphrases, or keyphrases as opposed to aspect specific opinion phrases which is the focus of this work. They tend to employ variations of term, segment, structural, syntactic, or rule/window based features as opposed to our proposed pivot features with respect to the head-aspect that allow modeling arbitrarily long opinion phrases. Also, above works that employ sequence modeling (e.g., [18, 20]), rely on canonical CRFs for phrase boundary detection. This has two shortcomings in our given problem context: (i) Canonical CRFs have a strong bias towards detecting a potential opinion phrase (issue) around the head aspect for every sentence in which the head aspect appears. This unfortunately leads to higher false positive rate as not every sentence mentioning the head aspect has an issue. We noticed this in our pilot studies, (ii) Under canonical CRFs, the space of potential issue phrases is much smaller than all possible enumerable sequences resulting in inaccurate phrase boundary detection. To address these shortcomings, we propose a two-step approach:

- Task I: Given a head aspect (HA), detect whether a sentence containing the HA mentions an issue.

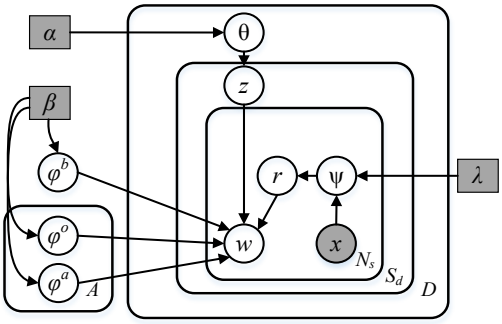

Figure 1: Plate notation of ME-ASM

- Task II: Given a HA and a sentence mentioning an issue, extract the issue phrase boundary.

To solve these tasks, we first posit a generative model, ME-ASM for domain-wise aspect specific sentiment extraction. ME-ASM provides us potential head aspects containing issues which directly feeds the issue annotation pipeline. Next, we use discriminative modeling for tasks I and II and leverage the generative model's posterior as features in the discriminative sequence model which significantly improves phrase extraction performance. To our knowledge, this has not been attempted before in opinion mining. The key contributions of this work include:

- A domain-wise aspect specific sentiment generative model for detecting head aspects.
- A family of pivot features for task I and phrase structural constraints for task II that can be used with generic discriminative and sequence modeling respectively.
- A comprehensive evaluation of the proposed methods against baselines including feature ablation and domain adaptation.
- A labeled data of aspect specific opinion phrases across 6 domains containing 3610 instances (sentences) tagged with phrase boundaries that imply negative opinions. The dataset used in this work is available at http://www.cicling.org/2016/data/10.

2 Aspect Specific Sentiment Modeling

We now present our generative semi-supervised model for extracting domain-wise aspect specific sentiments. As mentioned in §1, this is the first major step that feeds the pipeline for both tasks: [I] issue sentence classification, [II] issue phrase boundary extraction. Our model, ME-ASM (Max-Ent Aspect Sentiment Model) has resemblances to previous aspect extraction models [25, 26, 27, 28] but aims to deliver i) robust domain-wise aspects, ii) clear separation of aspects from aspects specific sentiments, and iii) sentence level modeling for sharper aspect extraction.

As noted in [25], modeling entire reviews as documents tend to correspond to the product's global properties (e.g., brand, name) resulting in overlapping aspects. To avoid this, we perform sentence level modeling. We posit $a_{1...A}$ aspects, $o_{1...A}$ aspect specific sentiments, and background language models using multinomials φ_a^A, φ_a^O, and

φ^b drawn from $Dir(\beta)$ respectively over the vocabulary $v_{1...V}$. For each domain d, we draw a domain specific aspect distribution $\theta_d \sim Dir(\alpha)$. Next, for each review sentence (document), s_d of a domain, d we draw an aspect, $z_{d,s} \sim Mult(\theta_d)$. We assume that each sentence evaluates one aspect which mostly holds in the review domain. Next, to generate each word, $w_{d,s,j}$ of the sentence, s_d we first set the switch variable $r_{d,s,j} \leftarrow Mult(\psi_{d,s,j})$ from a previously trained discriminative model (see §2.2). The switch variable, $r \in \{\hat{a}, \hat{o}, \hat{b}\}$ takes on values corresponding to aspect, sentiment, and background words as estimated by the switch model ψ. Finally, depending upon the latent aspect, $z_{d,s}$ and the switch variable $r_{d,s,j}$, we emit $w_{d,s,j}$ as follows:

$$w_{d,s,j} \sim \begin{cases} Mult(\varphi^b) & \text{if } r_{d,s,j} = \hat{b} \\ Mult\left(\varphi^A_{z_{d,s}}\right) & \text{if } r_{d,s,j} = \hat{a} \\ Mult\left(\varphi^O_{z_{d,s}}\right) & \text{if } r_{d,s,j} = \hat{o} \end{cases} \tag{1}$$

2.1 Inference

We employ MCMC Gibbs sampling for posterior inference. As latent variables z and r belong to different levels, we hierarchically sample z and then r for each sweep of a Gibbs iteration as follows:

$$p(z_{d,s} = a | Z_{\neg d,s}, R_{\neg d,s}, W_{\neg d,s}) \propto \frac{(n^s_{d,a})_{\neg d,s} + \alpha}{(n^s_{d,(\cdot)})_{\neg d,s} + A\alpha} \times \left[\left(\prod_{v=1}^{V} \frac{\Gamma(n^A_{a,v} + \beta)}{\Gamma\left(n^A_{a,v_{\neg d,s}} + \beta\right)}\right) / \left(\frac{\Gamma(n^A_{a,(\cdot)} + V\beta)}{\Gamma\left(n^A_{a,(\cdot)_{\neg d,s}} + V\beta\right)}\right)\right] \times \left[\left(\prod_{v=1}^{V} \frac{\Gamma(n^O_{a,v} + \beta)}{\Gamma\left(n^O_{a,v_{\neg d,s}} + \beta\right)}\right) / \left(\frac{\Gamma(n^O_{a,(\cdot)} + V\beta)}{\Gamma\left(n^O_{a,(\cdot)_{\neg d,s}} + V\beta\right)}\right)\right] \tag{2}$$

$$p(r_{d,s,j} = l | \ldots, w_{d,s,j} = v) \propto \frac{(n^l_{a,v})_{\neg d,s,j} + \beta}{(n^l_{a,(\cdot)})_{\neg d,s,j} + V\beta} \times \frac{\exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, l))}{\sum_{l \in \{\hat{a}, \hat{o}, \hat{b}\}} \exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,s,j}, l))}; \; l \in \{\hat{a}, \hat{o}, \hat{b}\} \tag{3}$$

where $n^s_{d,a}$ denotes the # of sentences in domain d assigned to aspect a. $n^A_{a,v}$, $n^O_{a,v}$, n^B_v denotes the # of times word v was assigned to aspect a in the aspect, aspect specific opinion, and background language models respectively. A count variable with subscript $(\cdot)$ signifies the marginalized sum over the latter index and $\neg$ denotes the discounted counts.

2.2 Setting Switch and Hyper-parameters

ME-ASM performs a three-way switch between aspects, sentiments and background words and is motivated by models in [27, 28]. We employ a discriminative Max-Ent model for performing the switch. As aspect and sentiment terms play different syntactic roles in a sentence, we leverage the part-of-speech (POS) and syntactic chunk tags of the terms as features for learning the Max-Ent model $\psi_{d,s,j}$ conditioned on the observed feature vector $\overrightarrow{x_{d,s,j}}$ associated with $w_{d,s,j}$. We use a window of 4 terms both ahead and behind the term $w_{d,s,j}$ to encode feature context. The Max-Ent λ were learned using 500 labeled terms in each domain following the approach in [28]. The hyper-parameters

Issue subject: Signal		Issue subject: Firmware	
Aspect (φ^A)	Sentiment (φ^O)	Aspect (φ^A)	Sentiment (φ^O)
signal	loses	firmware	bug
wireless	faded	hardware	upgrade
antenna	drops	*third*	update
wifi	poor	*party*	old
download	losing	version	restore
unsecured	unavailable	level	incompatible
feet	slow	driver	instable
router	*clear*	latest	*install*
range	weak	*download*	flashing

(a) Router Domain

Issue subject: Screen		Issue subject: Voice	
Aspect (φ^A)	Sentiment (φ^O)	Aspect (φ^A)	Sentiment (φ^O)
screen	small	voice	poor
touchscreen	sensitive	sound	clarity
display	unresponsive	*directions*	understand
touch	*bright*	accent	*sounds*
contrast	useless	command	quality
garmin	horrible	street	awful
resolution	*responsive*	name	*slow*
3d	clutter	instructions	horrible
map	*poor*	*english*	distorted

(b) GPS Domain

Table 1: Top ranked aspect and sentiment terms in two head aspects (issue subjects) across two domains. Clustering errors are *italicized in red*.

for ME-ASM were set as $\beta = 0.1$ and $\alpha = 50/A$ following [29]. The total # of aspects, A across all 6 domains (see §3.1) were set to 20 after tuning via our pilot experiments.

2.3 Estimated Posterior

Table 1 shows the top terms for the estimated posterior on φ_a^A and φ_a^O. Owing to space constraints, we focus on two aspects for two domains each. As our goal is to discover phrasal issues, we run our model on $\leq$ 3-star reviews (see dataset in §3.1). Except for some clustering errors (*italicized in red*), which is a known issue in generative modeling [30], we see that ME-ASM yields a decent clustering of aspects and aspect specific sentiment terms implying negative opinions. The posterior feeds head-aspect detection (§3.1) and tasks I and II.

3 Task I: Issue Sentence Classification

This section details the task I: Given a head aspect, HA and a sentence containing the HA, classify whether the sentence mentions an issue or not. We first detail our dataset, followed by features and results.

3.1 Dataset

To our knowledge, there are no publicly available datasets that mark phrase boundaries of aspect specific sentiment expressions. The closest tasks to ours are in (1) SemEval

2015 Aspect based Sentiment Analysis Task [31] which aims to discover opinion targets on entity-aspect pairs and have annotations such as {FOOD#QUALITY, "Chow fun", negative, from="0" to="8"} for the sentence: "Chow fun was dry; pork shu mai was…" where annotations apply to aspect expressions, and (2) The MPQA 2.0 corpus [2] although has some labeled opinion expressions, it mostly contains generic subjective expressions spanning dimensions such as sentiment, agreement, arguing, intention, etc. Both corpora don't contain the entire aspect specific sentiment phrase boundaries labeled and hence cannot be directly used in our task.

Hence, we constructed a data resource for the proposed task. Given our problem context, we consider 1, 2, and 3-star product reviews from Amazon.com. For each domain, we annotated issue phrases for top 4 aspects that had the highest appearance of negative opinions (estimated using the posterior on φ_a^A and φ_a^O from ME-ASM). We followed previous work in [32] for training our judges for annotation. A phrase was defined to be any subjective expression that captures various sentiments (evaluation, emotion, appraisal, etc.) toward the head aspect and containing the head aspect (see examples in §1). The annotation was distributed across four human judges (native English speakers). Every sentence was tagged by at least two judges. Inconsistencies were resolved by a third judge. Across each domain, we obtained, kappa $\kappa \in [0.71 - 0.82]$ indicating substantial to high agreements. The annotation statistics are reported below. For each domain, we report the head aspects and the counts as (x/y) where x is the # of issue sentences and y the total # of sentences in that domain: Router (1284/5063; connection, firmware, signal, wireless), GPS (632/2075; voice, software, screen direction), Keyboard (667/1446; spacebar, range, pad, keys), Mouse (494/2488; battery, button, pointer, wheel), MP3-Player (174/352; button, interface, jack, screen), Earphone (359/678; cord, jack wire). This dataset serves both of our tasks I and II.

3.2 Features

As product issues are directly reflected in the language usage, word and POS (W+POS) *n*-gram features serve as natural baselines. We consider unigrams and bigrams. Using (W+POS) features here is akin to traditional sentence polarity classification [33].

However, in our problem context, (W+POS) features are insufficient as they do not consider the head aspect and relevant positional/contextual features, i.e., how do different POS tags, syntactic units (chunks), polar sentiments appear in proximity to the head aspect? Hence, centering on the issue subject (head aspect), we propose a set of pivot features to model context.

Pivot Features: We consider five feature families which take on a set of values:

POS Tags (T): *DT*, *IN*, *JJ*, *MD*, *NN*, *RB*, *VB*, etc.
Phrase Chunk Tags (C): *ADJP*, *ADVP*, *NP*, *PP*, *VP*, etc.
Prefixes (P): *anti*, *in*, *mis*, *non*, *pre*, *sub*, *un*, etc.
Suffixes (S): *able*, *est*, *ful*, *ic*, *ing*, *ive*, *ness*, *ous*, etc.
Word Sentiment Polarity (W): *POS*, *NEG*, *NEU*

Pivoting on the head aspect, we look forward and backward to generate a family of binary features defined by a specific template (see Table 2). Each template generates several feature that capture various positional context around the head aspect. Additionally, we consider up to 3rd order pivot features allowing us to model a rich and expressive feature space.

Category	Feature Template	Example of feature appearing in a sentence
1st order features $X_{i+j}; -4 \leq j \leq 4$ $X \in \{T, C, P, S, W\}$	W_{i+j}	$W_{i-1} = NEG$; previous term of head aspect is of NEG polarity, … have this terrible *voice* on the…
	S_{i+j}	$S_{i-2} = ing$; suffix of 2nd previous term of head aspect is "ing", …kept dropping the *signal*…
	…	…
2nd order features $X_{i+j}, Y_{i+j}; -4 \leq j \leq 4$ $X, Y \in \{T, C, P, S, W\}$	$T_{i+j}, T_{i+j'}$	$T_{i-2} = JJ, T_{i-1} = VBZ$, …frequently drops *connection*…
	$T_{i+j}, C_{i+j'}$	$T_{i+2} = RB, C_{i+3} = ADJP$; …*screen* is too small…
	…	…
3rd order features $X_{i+j}, Y_{i+j}, Z_{i+j}$; $-4 \leq j \leq 4$ X, Y, Z $\in \{T, C, P, S, W\}$	$T_{i+j}, S_{i+j'}, T_{i+j''}$	$T_{i+2} = JJ, S_{i+4} = un, T_{i+4} = JJ$; …*screen* is blank and unresponsive…
	…	…

Table 2: Pivot Feature Templates. The subscript i denotes the position of the issue subject (HA) which is italicized and the subscript j denotes the position relative to i.

Latent Semantics (LS): The generative model yields us aspect (φ_a^A) and aspect specific sentiment terms (φ_a^O) for each head aspect, a. It also provides us the assignments of latent variables z and r in each sentence. We leverage this information by positing the following features: i) Top 50 terms of φ^A, φ^S, ii) # of words assigned to aspect, opinion, and background distributions in a sentence, i.e., $|r_{d,s,n} = a|$, $|r_{d,s,n} = o|$, $|r_{d,s,n} = b|$, iii) the aspect assignment for the sentence, $z_{d,s}$, iv) for each term $\{w|w \in \varphi_a^A, \varphi_a^O\}$, the signed positional index of w form the head aspect, a.

3.3 Results

We now evaluate the performance on the first task of issue sentence classification. Merging sentences of all head aspects per domain, we report classification results for each domain. Upon experimenting with various kernels (linear, RBF, sigmoid) and features selection schemes in our pilot, we finalized on a RBF kernel SVM [34] with $C = 10, \gamma = 0.01$ and χ^2 feature selection as our classifier as it performed best. Table 4 shows the 5-fold cross validation (CV) results across different feature sets. While inducing LS features, for each fold of 5-fold CV, ME-ASM was run on the full data excluding the test fold. The learned ME-ASM was then fitted to the test set sentences for generating the LS features of the test instances. We note the following observations:

- Across each domain, the pivot features significantly ($p < 0.01$) improve precision, recall, and F1 scores over the baseline features across all domains.
- LS features alone improve performance significantly ($p < 0.03$) and are close to combined W+POS+Pivot features' performance. Particularly, LS features improve

Feature Set	P	R	F1	Acc.
Word (W) + POS	68.8	59.6	63.9	82.8
W + POS + Pivot	74.1	66.4	70.0	85.5
Latent Semantics	72.9	64.7	68.5	84.9
All	81.5	69.6	75.1	88.2

(a) Router

Feature Set	P	R	F1	Acc.
Word (W) + POS	70.3	57.2	63.1	69.2
W + POS + Pivot	73.1	60.5	66.2	71.5
Latent Semantics	70.9	59.6	64.8	70.2
All	76.6	64.3	69.9	74.5

(b) Keyboard

Feature Set	P	R	F1	Acc.
Word (W) + POS	69.6	58.9	63.8	79.6
W + POS + Pivot	72.6	62.4	67.1	81.3
Latent Semantics	71.1	62.3	66.4	80.8
All	73.6	64.8	68.9	82.2

(c) GPS

Feature Set	P	R	F1	Acc.
Word (W) + POS	65.9	56.7	60.9	85.5
W + POS + Pivot	67.5	60.5	63.8	86.1
Latent Semantics	66.5	59.5	62.8	85.8
All	66.8	60.1	63.3	86.8

(d) Mouse

Table 3: (P)recision, (R)ecall, F1 scores, and (Acc)uracy in % of 5-fold CV for issue sentence classification per domain.

recall. We also note that although the LS feature space is smaller than pivot features, it can perform quite well, thereby indicating its discriminative strength.

- Across all domains and feature sets, we find that recall is relatively lower than precision. This is due to the rather myriad forms of implied issues (e.g., "small screen buttons," "firmware does not contain the fixes," "firmware would reboot itself," etc.) which are difficult cases. Nonetheless, we note that LS and pivot features significantly improve recall over W+POS features.
- Lastly, combining all feature sets yield the best performance with an average F1 of ≈ 0.69 and accuracy of ≈ 0.83 across a total of 3,077 test instances (issues sentences of 4 domains combined, see §3.1) spanning 4 domains showing that the issue sentence classification module can be fed to subsequent phrase sequence model as a pipelined model (§5.4).

4 Task II: Phrase Boundary Extraction

We now focus on task II: Given a HA and a sentence mentioning an issue, extract the issue phrase boundary. We consider a heuristic baseline and three tailored sequence models for this task.

4.1 Unsupervised Heuristic Baseline (UHB)

In this approach, we consider a rule based model. In our problem context, two cases arise:

- The opinion phrase is in between the head aspect and a negative sentiment constituting a part of a noun, verb, adjective or adverbial phrase (e.g., "*signal* was so weak," "loss of *connection*," etc.)

- The opinion phrase is spread out between the head aspect, a positive sentiment and a negator (e.g., "*signal* was not so strong," "couldn't get a stable *connection*")

For the first case, we extract the index of the head aspect, a proximal negative sentiment and emit the terms between them as the phrase. For the second, we sort the index of the head aspect, the proximal positive sentiment relative to the head aspect, and the proximal negator relative to the positive sentiment, and emit the phrase spanning the minimum to maximum index. We consider a 5 term window for our proximity measure (tuned via pilot experiments) and use the associated[1] sentiment lexicon of [1]. This method serves as our baseline. Although heuristic, we will see that it can discover relevant opinion phrases.

4.2 Sequence Modeling

Recall from §1 that issue extraction requires us to detect a sequence of words (phrase) that directly or indirectly implies an aspect specific opinion. Let $\boldsymbol{x} = (x_1, \ldots x_n)$ denote the sequence of observed words in a sentence, and let each observation x_i has a label $y_i \in Y$ indicating whether x_i is part of an issue phrase, where $Y = \{B, I, O\}$. The state space of labels follow the standard BIO notation as described in [35], where values taken by y_i denote the *B*egin, *I*nside and *O*utside phrase alignments. Given a sentence containing an issue, $\boldsymbol{x}$**,** the extraction task is to find the best label sequence $\hat{y}$ that describes an issue. We employ a first order Markov linear-chain CRF whose predictor takes the following form,

$$p(y|x, \Lambda) = \frac{\exp(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x))}{Z(x, \Lambda)} \tag{4}$$

where f_k denotes the feature functions, $\Lambda = \{\lambda_k\}$ denotes the feature weights, and $Z(x, \Lambda)$ the normalization constant which takes the following form,

$$Z(x, \Lambda) = \sum_{y'} \big(\exp\big(\sum_i \sum_k \lambda_k f_k(y'_{i-1}, y'_i, x)\big)\big) \tag{5}$$

Given a set of training examples $\{\boldsymbol{x_j}, \overline{\boldsymbol{y_j}}\}$ where $\overline{\boldsymbol{y_j}}$ are the correct labels, we estimate the parameters by minimizing the negative log-likelihood (NLL),

$$\Lambda = \underset{\Lambda}{\operatorname{argmin}} \Big(-\sum_j \log\Big(p(\overline{y_j}|x_j, \Lambda)\Big) + \sum_k \lambda_k^2\Big) \tag{6}$$

The term $\sum_k \lambda_k^2$ indicates L_2 regularization on the feature weights, λ_k. It penalizes the NLL to prevent extreme values for λ_k. We experiment with both CRF and CRF with L_2 regularization (CRF-L2R).

4.3 Linear Chain Features

We now describe the encoding of our linear chain features (LCF), $f(y_{i-1}, y_i, x)$. We use the templates of the pivot features (Table 2) with a few changes. The index i for LCF in templates refers to the (current) position of any word in the sentence and not necessarily the head aspect. Further, in addition to the families defined under pivot features (§3.2), we consider latent semantic (LS) features (φ_a^A and φ_a^O for each head aspect, a). These can be very useful as encoding the position of the aspect specific senti-

[1] http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

ment terms (see Table 1) relative to the head aspect can guide phrase boundary detection. Further, each feature generated by the above defined templates is coupled with the value of the current label y_i and a combination of current and previous labels y_i, y_{i-1}.

4.4 Phrase Structural Constraints

Although the above canonical CRF formulation can detect phrase boundaries, it has one key downside. Under canonical CRFs, the predictor models a large probability space (the denominator, $Z(x, \Lambda)$ in (5)) as it considers all possible sequence labelings. This unfortunately results in sparse probabilities for potential issue phrases. This is so because all issue sequences, $\boldsymbol{y} = \{y_i\}$ always exhibit the pattern that exactly one component of $\boldsymbol{y}$, say $y_l = B$ followed by one or more consecutive I, ($y_{>l} = I$), followed by all O. So only candidate phrases that conform to the above pattern are *valid* phrases for issues. Motivated by previous work in [36], we consider a constrained model where $Z(x, \Lambda)$ is reduced to only *valid* patterns. This reduces the contention of incorrect sequences thereby assuaging the sparsity problem. The key difference lies in summing over only valid sequences in (5). We employ L_2 regularization and refer it by CRF-PSC.

5 Evaluation

This section details results for sentiment phrase extraction (Task II) (§4).

5.1 Qualitative Analysis

Table 4 shows the sample phrases extracted by UHB and CRF for two domains. Here we report the base CRF model (and not CRF-PSC) as it is representative of other CRF extensions. However, all four models are compared quantitatively in the subsequent sections. From Table 4, we note that although UHB is rule based, it can discover some phrases correctly including harder/longer phrases ("slow down or drop connection"). However, owing to non-reverence to sequential structure, it has two key downsides: i) incorrect grammatical structures, ii) incoherent phrase extraction (i.e., does not capture the key issue). These are overcome by CRF's sequence modeling.

5.2 Quantitative Results

Here we focus on per domain cross aspect analysis and also compare models across different domains. For each domain (see §3.1), we test on one head aspect by applying the sequence model learned from examples of the rest 3 head aspects for that domain. This gives us one set of results for one head aspect in a domain. Repeating it for other head aspects of that domain and averaging the performances over all head aspects for a domain, allows us to estimate the comprehensive performance of a model on a given domain. This is akin to 4-fold cross validation per aspect for each domain. Also, for inducing generative LS features in CRF models, we follow the same technique as used in task I (§3.3). We use the standard token overlap metric for evaluating the phrase

Head Aspect	UHB	CRF
signal	signal was so weak *good signal sometimes I don't* signal losing *signal it can interfere* problems sending signal *nothing to improve signal*	stopped broadcasting a signal starts dropping signal has a weak signal frequently loses signal signal faded very quickly *weak signal like before*
connection	connection drop problems loss of connection *connection or a very poor* slow down or drop connection *connection refused* connection dies	connection would break constantly drops connection drop your connection connection would drop out *hold connection steady for* connections don't last

(a) Router Domain

Head Aspect	UHB	CRF
screen	screen software crashes *defective but the screen* *lack of a screen* screen went dead *damaging the screen* *screen has odd*	*screen doesn't come on* screen started to fade screen is too small screen went black *screen doesn't come* screen will go black
voice	voice is very distorted *voice files and unneeded* *voice doesn't disturb* *disable the voice* voice prompts were slow *interrupt its voice*	voice has a certain grating voice is very distorted *delete the foreign voices* voice is very shaky voice is scratchy voice is pretty feeble

(b) GPS Domain

Table 4: Qualitative comparison of aspect specific opinion phrases discovered by two methods (a) Unsupervised Heuristic Baseline (UHB), vs. (b) Linear chain CRF for two head aspects each across two domains. Extraction errors are *italicized in red*.

boundaries. For each sentence $s \in S$, if s_c and s_p denote the correct and predicted expression spans (tokens) of the target phrase in s, then $recall(r) = \underset{s \in S, |s_c| \neq 0}{\text{avg}} \left(\frac{|s_c \cap s_p|}{|s_c|} \right)$, $precison(p) = \underset{s \in S, |s_p| \neq 0}{\text{avg}} \left(\frac{|s_c \cap s_p|}{|s_p|} \right)$, and $F = \frac{2pr}{p+r}$ where S is the set of sentences in the test fold of cross-validation. From Table 5, we note the following observations:

- Across each domain, we note that F1 scores of sequence models progressively improve in the following performance order CRF→CRF-L2R→CRF-PSC over the rule based baseline, UHB. This shows sequence modeling is useful in detecting issue phrases.
- Gains in F1 score of CRF, CRF-L2R over UHB are significant (see Table 5 caption). CRF-PSC further improves the result (especially recall) showing that encoding phrase structural constraints in sequence modeling for this task is useful.
- GPS and Mouse domains seem harder as the performance of all models are relatively lower than Keyboard and Router. This can be linked with the relative amount of training examples for each domain (see §3.1).

Model	P	R	F1
UHB	67.0	73.7	70.2
CRF	87.9	76.9	82.0
CRF-L2R	88.7	77.1	82.5
CRF-PSC	92.0	80.1	85.7

(a) Router

Model	P	R	F1
UHB	64.4	68.1	66.2
CRF	86.8	74.8	80.3
CRF-L2R	87.6	75.7	81.2
CRF-PSC	90.4	78.3	83.9

(b) Keyboard

Model	P	R	F1
UHB	67.5	67.7	67.6
CRF	84.1	71.9	77.5
CRF-L2R	84.8	72.7	78.3
CRF-PSC	86.7	73.5	79.5

(c) GPS

Model	P	R	F1
UHB	60.7	59.1	59.8
CRF	83.8	61.6	70.9
CRF-L2R	84.5	61.9	71.4
CRF-PSC	86.1	63.6	73.1

(d) Mouse

Table 5: Precision, recall, F1 scores of sequence models on phrase boundary detection. Gains of CRF and CRF-L2R over UHB are significant at $p < 0.01$. Gains of CRF-PSC over CRF-L2R are significant at $p < 0.02$. Significance was measured using *t*-test across all domains.

5.3 Feature Ablation

In order to assess the relative discriminative strengths of various feature families, we perform ablation analysis. We fix our model to CRF-PSC (as it performed best) and use the F1 metric. Starting from the full feature set, we drop each feature family and report the resulting performance. From Table 6, we note that each feature family has a positive contribution toward the phrase extraction task as dropping it has a statistically significant ($p < 0.05$) reduction in F1 across each domain. Dropping Latent Semantic (LS) features, POS Tags and word polarity impacts performance substantially. Especially, the LS feature family as the LS features help locate the index of the aspect specific sentiment terms in a sentence that guides the issue phrase boundary detection. Thus, we can see that all feature families are useful, with some (e.g., Latent semantics) contributing significantly to phrase boundary extraction performance.

5.4 Domain Adaptation

We now consider a realistic setting of applying our pipelined model to two new domains: Earphone and MP3 Players. We first train the issue sentence classifier and phrase extraction sequence models on other 4 domains. Then the issue sentences identified by the first (classification) model are fed to the previously trained sequence model (CRF-PSC) for phrase extraction on those sentences. We consider two systems and report intermediate results (prec., rec., F1, acc.) of issue sentence classification and also phrase boundary extraction in Table 7. The precision, recall, and F1 for phrase extraction were computed as defined in §5.2 whereby losses in classification task (e.g., false negative/false positive issue sentences) are accounted (as a penalty) in the recall/precision for phrase extraction respectively.

We note relatively higher precision and recall in classification performance (col 3, 4,

Dropped Feature	Router	Keyboard	GPS	Mouse
None	85.7	83.9	79.5	73.1
POS Tags	**81.7**	**80.1**	**75.8**	**70.0**
Phrase Chunk Tags	83.8	81.7	77.3	71.3
Word Prefix/Suffix	84.1	82.5	78.1	72.4
Word Sent. Pol.	82.7	81.8	77.6	71.7
Latent Semantics	**80.1**	**77.9**	**74.6**	**68.7**

Table 6: F1 scores of CRF-PSC upon feature ablation.

System	Domain	Prec	Rec.	F1	Acc	P-Seq	R-Seq	F1-Seq
SVM (W+POS)+ UHB	Earphone	80.2	70.6	75.1	75.2	53.7	53.0	53.3
	MP3 player	84.1	73.5	78.4	80.0	56.3	55.1	55.7
	Avg.	82.2	72.1	76.8	77.6	55.0	54.1	54.5
SVM (W+POS+ LS+Pivot) + CRF-PSC	Earphone	84.7†	76.7†	80.5†	80.3	75.3	62.9	68.6
	MP3 player	88.6†	79.8†	83.9†	84.9	78.9	65.5	71.5
	Avg.	**86.7**	**78.3**	**82.2**	**82.6**	**77.1**	**64.2**	**70.1**

Table 7: Pipeline model results. Prec., Rec., F1, Acc apply to issue sentence classification. P-Seq, R-Seq, and F1-Seq apply to performance of phrase extraction on the target domain. † indicates significance at $p < 0.01$ measured via t-test.

Table 7) compared to individual domain experiments (Table 3) as now we have more training data (combination of 4 domains – Router, Keyboard, Mouse, GPS). Both systems (col 1, Table 7) find the Earphone domain harder than the MP3-Player domain. One possible reason for this could be that the domain of MP3 Players shares common aspects (e.g., screen, button) with training domains Mouse, GPS that help improve knowledge transfer. Nonetheless, on average, our system SVM (W+POS+LS+Pivot)+CRF-PSC significantly outperforms the baseline, SVM(W+POS)+UHB yielding an average F1 score of 82.2% for task I and 70.1% for task II showing decent generalization performance across new domains.

6 Conclusion

This work performed an in-depth analysis of a novel task in sentiment analysis – aspect specific opinion phrase extraction. The paper focused on phrases implying negative opinions (*issues*) in the product reviews domain. First, a generative model, ME-ASM was employed for discovering the top head aspects in each domain having potential issues. Next, the sentences containing the head aspect (issue subjects) were annotated for issues including issue phrase boundaries. Discriminative large-margin and sequence models using pivot features were employed to classify issue sentences and extract issues phrase boundaries respectively. Experimental results showed that the proposed approach outperformed baseline systems and also facilitated inductive knowledge transfer across domains. The paper also contributes a new large resource of labeled aspect specific sentiment expressions across 6 domains that can serve for various sequence modeling researches/tasks in opinion mining.

References

1. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proc. 2004 ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '04. ACM Press, New York, New York, USA, p 168
2. Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Lang Resour Eval 39:165–210.
3. Wang X, McCallum A, Wei X (2007) Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Data Mining, 2007. ICDM 2007. Seventh IEEE Int. Conf. pp 697–702
4. Fei G, Chen Z, Liu B (2014) Review Topic Discovery with Phrases using the P{ó}lya Urn Model. COLING
5. Breck E, Choi Y, Cardie C (2007) Identifying expressions of opinion in context. Proc 20th Int Jt Conf Artifical Intell 2683–2688.
6. Choi Y, Cardie C, Riloff E, Patwardhan S (2005) Identifying sources of opinions with conditional random fields and extraction patterns. In: Proc. Conf. Hum. Lang. Technol. Empir. Methods Nat. Lang. Process. - HLT '05. Association for Computational Linguistics, Morristown, NJ, USA, pp 355–362
7. Choi Y, Breck E, Cardie C (2006) Joint extraction of entities and relations for opinion recognition. Proc 2006 Conf Empir Methods Nat Lang Process 431–439.
8. Li H, Mukherjee A, Liu B, Si J (2015) Extracting Verb Expressions Implying Negative Opinions. Proc. twenty-ninth AAAI Conf. Artif. Intell.
9. Berend G (2011) Opinion Expression Mining by Exploiting Keyphrase Extraction. In: Int. Jt. Conf. Nat. Lang. Process. pp 1162–1170
10. Johansson R, Moschitti A (2010) Reranking models in fine-grained opinion analysis. Proc 23rd Int Conf Comput Linguist 519–527.
11. Yang B, Cardie C (2012) Extracting opinion expressions with semi-markov conditional random fields. In: Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. pp 1335–1345
12. Klinger R, Cimiano P (2013) Bidirectional Inter-dependencies of Subjective Expressions and Targets and their Value for a Joint Model. Assoc. Comput. Linguist. (Short Pap.
13. Kim S-M, Hovy E (2006) Extracting opinions, opinion holders, and topics expressed in online news media text. Proc Work Sentim Subj Text, Assoc Comput Linguist 1–8.
14. Wu Y, Zhang Q, Huang X, Wu L (2009) Phrase dependency parsing for opinion mining. Proc 2009 Conf Empir Methods Nat Lang Process 1533–1541.
15. Jakob N, Gurevych I (2010) Extracting opinion targets in a single- and cross-domain setting with conditional random fields. Proc 2010 Conf Empir Methods Nat Lang Process 1035–1045.
16. Kobayashi N, Inui K, Matsumoto Y (2007) Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. Proc. 2007 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. (EMNLP-CoNLL
17. Johansson R, Moschitti A (2011) Extracting opinion expressions and their polarities: exploration of pipelines and joint models. Assoc Comput Linguist (Short Pap 101–106.
18. Yang B, Cardie C (2014) Joint Modeling of Opinion Expression Extraction and Attribute Classification. Trans Assoc Comput Linguist 2:505–516.

19. Sauper C, Haghighi A, Barzilay R (2011) Content models with attitude. Proc 49th Annu Meet Assoc Comput Linguist Hum Lang Technol - Vol 1 350–358.
20. Yang B, Cardie C (2013) Joint Inference for Fine-grained Opinion Extraction. In: Assoc. Comput. Linguist. pp 1640–1649
21. Barzilay R, McKeown KR (2001) Extracting paraphrases from a parallel corpus. In: Proc. 39th Annu. Meet. Assoc. Comput. Linguist. pp 50–57
22. Apidianaki M, Verzeni E, McCarthy D (2014) Semantic Clustering of Pivot Paraphrases. In: Conf. Lang. Resour. Eval. pp 4270–4275
23. Keshtkar F, Inkpen D (2010) A corpus-based method for extracting paraphrases of emotion terms. In: Proc. NAACL HLT 2010 Work. Comput. approaches to Anal. Gener. Emot. Text. pp 35–44
24. Hasan KS, Ng V (2014) Automatic keyphrase extraction: A survey of the state of the art. Proc 52nd Annu Meet Assoc Comput Linguist 1262–1273.
25. Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. In: Proc. 17th Int. Conf. World Wide Web. pp 111–120
26. Brody S, Elhadad N (2010) An unsupervised aspect-sentiment model for online reviews. Annu Conf North Am Chapter Assoc Comput Linguist 804–812.
27. Mukherjee A, Liu B (2012) Aspect Extraction through Semi-Supervised Modeling. Assoc. Comput. Linguist.
28. Zhao WX, Jiang J, Yan H, Li X (2010) Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proc. 2010 Conf. Empir. Methods Nat. Lang. Process. pp 56–65
29. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci U S A 101:5228–5235.
30. Chang J, Gerrish S, Wang C, Boyd-graber JL, Blei DM (2009) Reading tea leaves: How humans interpret topic models. In: Adv. Neural Inf. Process. Syst. pp 288–296
31. Pontiki, M., Galanis, D., Papageogiou, H., Manandhar, S., & Androutsopoulos I (2015) Semeval-2015 task 12: Aspect based sentiment analysis. Proc. 9th Int. Work. Semant. Eval. (SemEval 2015), Denver, Color.
32. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proc. Conf. Hum. Lang. Technol. Empir. Methods Nat. Lang. Process. - HLT '05. Association for Computational Linguistics, Morristown, NJ, USA, pp 347–354
33. Yu H, Hatzivassiloglou V (2003) Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Proc. 2003 Conf. Empir. methods Nat. Lang. Process. pp 129–136
34. Joachims T (1999) Making large-scale support vector machine learning practical. Adv. Kernel Methods
35. Ramshaw LA, Marcus MP (1995) Text chunking using transformation-based learning. ACL Third Work. Very Large Corpora
36. Li Y, Jiang J, Chieu HL, Chai KMA (2011) Extracting Relation Descriptors with Conditional Random Fields. In: Int. Jt. Conf. Nat. Lang. Process. pp 392–400