Extracting Aspect Specific Opinion Expressions

Abhishek Laddha
IIT Delhi *
New Delhi, India, 110016
laddhaabhishek11@gmail.com

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

Abstract

Opinionated expression extraction is a central problem in fine-grained sentiment analysis. Most existing works focus on either generic subjective expression or aspect expression extraction. However, in opinion mining, it is often desirable to mine the aspect specific opinion expressions (or aspect-sentiment phrases) containing both the aspect and the opinion. This paper proposes a hybrid generative-discriminative framework for extracting such expressions. The hybrid model consists of (i) an unsupervised generative component for modeling the semantic coherence of terms (words/phrases) based on their collocations across different documents, and (ii) a supervised discriminative sequence modeling component for opinion phrase extraction. Experimental results using Amazon.com reviews demonstrate the effectiveness of the approach that significantly outperforms several state-of-the-art baselines.

1 Introduction

Aspect based sentiment analysis is one of the main frameworks in opinion mining (Liu and Zhang, 2012). Most of the websites only display the aggregated ratings of products but people are more interested in fine-grained opinions that capture aspect specific properties in reviews. Therefore, it is desirable to have a holistic approach to mine aspect specific opinion expressions containing both aspect and opinion terms within the sentence context as a composite aspect-sentiment phrase (e.g., "had to flash firmware everyday", "clear directions in voice" etc.) and further group them under coherent aspect categories. Apart from knowing the key issues in products that are often expressed via aspect-sentiment phrases, they are also useful in applications such as comparing similar products and summarizing their important features where it is more convenient to have the aspect-sentiment phrases rather than generic aspect/sentiment words lacking the natural aspect opinion correspondence in the right context. They can also be applied to the various tasks such as sentiment classification, comparative aspect evaluations, aspect rating prediction, etc.

The thread of research in (Brody and Elhadad, 2010; Titov and McDonald, 2008; Zhao et al., 2010; Mei et al., 2007; Jo and Oh, 2011) focus on extracting and grouping aspect and opinion words via generative models but lack the natural aspect opinion correspondence (e.g., in the manner they appear in sentences). (Wang et al., 2016; Fei et al., 2016) can discover aspect specific opinion unigrams but does not focus on phrases. The thread on fine grained opinion expressions (Wiebe et al., 2005; Choi et al., 2006; Breck et al., 2007) focus on subjective expression extraction which are generic instead of aspect specific. Formally, the task can be stated as follows:

Given a set of reviews, for each sentence, $s = (w_1, \ldots w_n)$, with the head aspect $(HA), w_{HA=i}, i \in [1, n]$, discover a sub-sequence $(w_p, \ldots w_q)$ where $p \leq i \leq q$ that best describes the aspect-sentiment phrase containing the head aspect. We refer head aspect to the word describing

Research performed during author's internship at University of Houston

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 627–637,
Austin, Texas, November 1-5, 2016.

fine-grained property of product. Further, group these phrases under relevant aspect categories. The examples below show labeled aspect-sentiment phrases within [[ ]] with the head aspect (HA) italicized:

- I've been very happy with it so far done a [[*firmware* update without a hitch]].
- After less than two years, the [[*signal* became spotty]].

In this paper, we propose a novel hybrid model to solve the problem. We call this Phrase Sentiment Model (PSM). PSM is capable of extracting a myriad of expression types covering: verb phrases ("screen has poor viewability"), noun, adjective or adverbial phrase ("recurrent black screen of death", "quite stable and fast connection"), implied positive ("voice activated directions"), implied negative ("requires reboot every few hours") etc. The hybrid framework facilitates holistic modeling that caters for varied expression types (leveraging its discriminative sequence model) and also grouping them under relevant aspect categories with context (exploiting its generative framework). Our approach is also context and polarity independent facilitating generic aspect-sentiment phrase extraction in any domain.

Further, we propose a novel sampling scheme based on Generalized Pólya urn models that optimizes phrasal collocations to improve coherence. To the best of our knowledge, a hybrid framework has not been attempted before for opinion phrase extraction. Additionally, the paper produced a labeled dataset of aspect specific opinion phrases across 4 domains containing more than 5200 sentences coded with phrase boundaries across both positive and negative polarities which will be released to serve as a language resource. Experimental evaluation shows that our approach outperformed the baselines by a large margin.

2 Related Work

Subjective expression extraction (Choi et al., 2005) has traditionally used sequence models (e.g., CRFs). Various parsing, syntactic, lexical and dictionary based features (Kim and Hovy, 2006; Jakob and Gurevych, 2010; Kobayashi et al., 2007) have been used for subjective expression extraction. In (Yang and Cardie, 2013; Johansson and Moschitti, 2011) dependency relations were also used for opinion expression extraction. Sauper et al., (2011) employs an HMM over words and model the latent topics as states in an HMM to discover the product properties (often aspects) and its associated attributes (pos/neg) polarities separately. In Yang and Cardie, (2012) a semi-CRF based approach is used which allow sequence labeling at segment level and (Yang and Cardie, 2014) employed semi-CRF for opinion expression intensity and polarity classification. However, all the above works focus on generic subjective expressions as opposed to aspect specific opinion-sentiment phrases.

In (Choi et al., 2006; Yang and Cardie, 2013) joint models were proposed for identifying opinion holders and expression, relations among them in news articles. In (Johansson and Moschitti, 2011) a re-ranking approach was used on the output of a sequence model to improve opinion expression extraction. In (Li et al., 2015; Mukherjee, 2016) subjective expressions implying a negative opinion were discovered using sequence models and markov networks; while in (Berend, 2011) supervised keyphrase extraction was used for phrase extraction. These works mostly relied on word level features under the first-order Markov assumption. Above works are tailored for only expression extraction and do not group coherent phrases under relevant aspect categories.

Another thread of research involves topic phrase mining. Wang et al., (2007) proposes a Topical n-gram model (TNG) that mines phrases based on statistical collocation. Lindsey et al., (2012) employ hierarchical Pitman-Yor process to model phrases. In Fei et al., (2014), Generalized Pólya urn model (LDA-P-GPU) was used to group the candidate noun phrases. In (El-Kishky et al., 2014; Liu et al., 2015), frequency based information were used for mining phrases that are good for generic phrases but cannot model relevant yet longer phrases due to their infrequency. Thus, they are unable to capture long phrases containing both aspect and opinion. The models TNG and LDA-P-GPU are closest to our task as they can discover relevant aspect expressions that can contain opinions and are considered as baselines.

Next there are works that generate phrasal

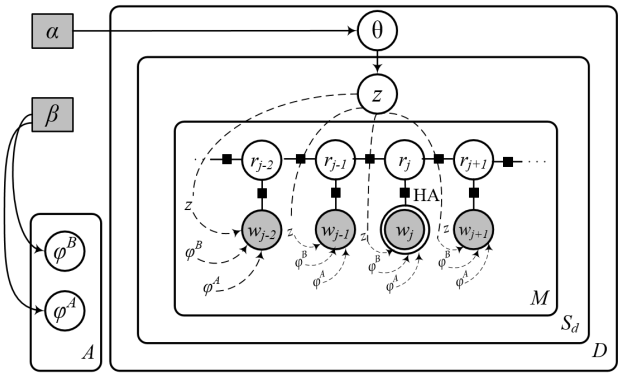

Figure 1: Plate Notation of PSM

datasets. In SemEval 2015, Aspect based Sentiment Analysis Task (Pontiki et al., 2015), a dataset was produced that had annotations for aspect phrases. The focus was on aspect phrases as opposed to aspect-sentiment phrases. The MPQA 2.0 corpus (Wiebe et al., 2005) has some labeled opinion expressions, but they are generic subjective expressions as opposed to aspect-sentiment phrases (see Section 1) we find in reviews.

Zhao et al., (2011) extracts topical phrase in tweets using relevance and interestingness. Wu et al., (2009) proposed a phrase dependency parsing approach to extract product feature (aspect expression) and opinion expression and the relation between them. They considered all noun and verb phrases (NPs, VPs) as product features and its surrounding dictionary opinion words as opinions. Features were constructed using phrase dependency tree to extract relation among all product features and opinions that were later used in aspect and opinion expression extraction. Although, Wu et al., (2009) doesn't discover aspect specific opinion phrases, its use of NPs in extracting candidate opinion phrases is similar to Fei et al., (2014) which is considered as a baseline.

3 Phrase Sentiment Model (PSM)

PSM is a hybrid between generative and discriminative modeling that combines the best of both worlds. Its generative modeling lays the foundation for emission of aspects and aspect specific opinion phrases in documents, while its discriminative sequence modeling component (via an embedded CRF) facilitates aspect specific opinion phrase extraction.

As noted in Titov and McDonald (2008), modeling entire reviews as documents tend to correspond to the global properties of a product (e.g., brand, name, etc.) resulting in rather overlapping aspects. To avoid this, we perform sentence level modeling that helps improve aspect sharpness. A review sentence s_d of N words, is denoted as $w_{d,s,j} = \{w_{d,s,1}, w_{d,s,2} \dots w_{d,s,N}\}$ where each $w_{d,s,j}$ is one of the V words in the vocabulary. There can be exponential number of phrase sequence possible for each sentence s_d i.e. $\langle w_{d,s,j} \rangle_{j=st}^{j=st+len-1}$ of arbitrary length $len (len = 1$ for words; $len > 1$ for phrases) starting at an index $st \in [0, |s_d|]$. We observed that most of opinion expression are centered around the head aspect thereby causing the space of potential opinion expressions to be quite sparse. We took advantage of following observation and trained a sequence model (e.g., CRF) for phrase sequence tagging as described in the next subsection. We generated $M = 5$ best sequence labelings of each sentence $s_d (s_d^{m \in 1 \dots M})$ via forward Viterbi and backward A^* search[1]. Hence, our vocabulary is the union of unigrams and n-grams discovered by CRF over M best sequence labeling, i.e., the model's vocabulary, $V = \{w_{d,s,j}\} \cup \{\langle w_{d,s_d^m,j} \rangle_{j=st}^{j=st+len-1}\} \; \forall d, s, j, len, st, m.$

In PSM, for each aspect a, we model its aspect specific terms (words/phrases) distributions and aspect background word distributions using multinomials φ_a^A and φ_a^B, drawn from $Dir(\beta)$ over the vocabulary $v_{1 \dots V}$. For each domain d, we first 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 true in the review domain. We associate a latent switch variable for each word $r_{d,s,j}$ and each phrase $\langle r_{d,s_d^m,j} \rangle_{j=st}^{j=st+len-1}$ of vocabulary where each switch variable $r \in \{0, 1\}$. To generate each term $\langle w_{d,s_d^m,j} \rangle_{j=st}^{j=st+len-1}$ of the labeled sequence $s_d^{m \in \{1 \dots M\}}$, we first set the switch variables for the sentence, s_d via the discriminative CRF model, i.e. $\langle r_{d,s_d^m,j} \rangle_{j=1}^{j=|s_d|} \leftarrow \frac{1}{Z} exp\left(\sum_j \sum_k \lambda_k f_k \left(r_{j-1}, r_j, w_j\right)\right)$ by fitting a previously trained CRF model. The switch variables, $r \in \{1, 0\}$ for a particular tagging s_d^m of

[1]Value of M was tuned via pilot experiments using the CRF++ toolkit (Kudo, 2009)

sentence s_d span over all its words and take values $r = 1$ for words being part of an aspect specific opinion phrase or $r = 0$ for aspect background words, upon observing all words in s_d. Finally, depending upon the aspect, $z_{d,s}$ and the switch variable $r_{d,s,j}$, we emit (unigram) terms in the sentence as follows:

$$w_{d,s,j} \sim \begin{cases} Mult(\varphi^A_{z_{d,s}}) & \text{if } r_{d,s,j} = 1 \\ Mult(\varphi^B_{z_{d,s}}) & \text{if } r_{d,s,j} = 0 \end{cases} \quad (1)$$

and for phrasal terms (i.e., when $\langle r_{d,s^m_d,j}\rangle_{j=1}^{j=st+len-1} = 1 \forall$ valid st and len), we emit $\langle w_{d,s^m_d,j}\rangle_{j=st}^{j=st+len-1} \sim Mult(\varphi^A_{z_{d,s}})$.

3.1 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:

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

Samplers for r consist of three cases: (i) individual aspect-specific opinion words, (ii) individual background words, and (iii) phrasal opinion:

$$p(r_{d,s,j} = 1|z_{d,s} = a, w_{d,s,j} = v, \ldots) \propto \frac{(n^A_{a,v})_{\neg d,s,j}+\beta)}{(n^A_{a,(\cdot)})_{\neg d,s,j}+V\beta)} \times p_{CRF}(r_{d,s,j-1}, r_{d,s,j} = 1|v) \quad (3)$$

$$p(r_{d,s,j} = 0|z_{d,s} = a, w_{d,s,j} = v, \ldots) \propto \frac{(n^B_{a,v})_{\neg d,s,j}+\beta)}{(n^B_{a,(\cdot)})_{\neg d,s,j}+V\beta)} \times p_{CRF}(r_{d,s,j-1}, r_{d,s,j} = 0|v) \quad (4)$$

$$p(\langle r_{d,s,j} = 1\rangle_{j=st}^{j=st+len-1}|z_{d,s} = a, \langle w_{d,s,j} = v'\rangle_{j=st}^{j=st+len-1}, \ldots) \propto \frac{(n^A_{a,v})_{\neg d,s,j}+\beta)}{(n^A_{a,(\cdot)})_{\neg d,s,j}+V\beta)} \times p_{CRF}(r_{d,s,j=st-1}, r_{d,s,j=st} = 1, r_{d,s,j=st+1} = 1, \ldots r_{d,s,j=st+len-1} = 1|v') \quad (5)$$

where $n^s_{d,a}$ denotes the # of sentences in domain d assigned to aspect a, $n^A_{a,v}, n^B_{a,v}$ denote the # of times term v was assigned to aspect a in the aspect specific opinion, and aspect specific background language models respectively. A count variable with subscript $(\cdot)$ signifies the marginalized sum over the latter index and $\neg$ denotes the discounted counts. The sampler in (5) computes the likelihood of a sequence of contiguous terms, $\langle r_{d,s,j} = 1\rangle_{j=st}^{j=st+len-1}$ forming an aspect ($z_{d,s} = a$) specific opinion phrase, v' starting at index $j = st$ and of length len, where $v' = \langle w_{d,s,j}\rangle_{j=st}^{j=st+len-1}$. The sequence probabilities, p_{CRF} in equations (3, 4, 5) can be obtained as follows.

Let $\mathbf{w} = (w_{t=1} \ldots w_{t=T})$ denote the sequence of observed words in a sentence, and let each observation w_t have a label $y_t \in Y$ indicating whether w_t is part of a aspect specific opinion phrase, where $Y = \{1, 0\}$. We consider a first order Markov linear-chain CRF in our hybrid model. We define the Markovian transition and forward-backward variables of our embedded CRF as follows:

$$\psi_t(j, i, w) = p(y_t = j|y_{t-1} = i)p(w_t = w|y_t = j) \quad (6)$$

$$\alpha_t(j) = \textstyle\sum_{i\in Y} \psi_t(j, i, w_t)\alpha_{t-1}(i) \quad (7)$$

$$\beta_t(j) = \textstyle\sum_{j\in Y} \psi_{t+1}(j, i, w_{t+1})\beta_{t+1}(j) \quad (8)$$

where $\alpha_1(j) = \psi_1(j, y_0, w_1)$ and $\beta_T(i) = 1$. This lays the foundation for expressing the sequence probabilities, p_{CRF} in closed form as follows:

$$p_{CRF}(y_{t-1}, y_t|w) \propto \alpha_{t-1}(y_{t-1})\psi_t(y_t, y_{t-1}, w_t)\beta_t(y_t) \quad (9)$$

$$p_{CRF}(y_{t-2}, y_{t-1}, y_t|w) \propto \alpha_{t-2}(y_{t-2})\psi_{t-1}(y_{t-1}, y_{t-2}, w_{t-1})\psi_t(y_t, y_{t-1}, w_t)\beta_t(y_t) \quad (10)$$

Eq. (9) is used for computing the sequence probabilities for individual opinion/background words for samplers in eq. (3, 4) while eq. (10) and its extensions are used for computing the sequence probabilities in the phrase samplers in eq. (5). The values for $\psi_t, \alpha_t, \beta_t$ are obtained from a previously trained CRF model upon fitting to the current sentence s_d for which sampling is being performed.

3.2 Embedded CRF Training

We employ linear-chain CRFs (Lafferty et al., 2001) for modeling phrases. While word (W) and Part-Of-Speech (POS) tag features are effective in various sequence modeling tasks (Yang and Cardie, 2014; Yang and Cardie, 2012), in our problem context, (W+POS) features are insufficient as they do not consider the head aspect (HA) and its 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 HA, we propose a set of pivot features to model context.

Pivot Features: We consider five feature families:
POS Tags (T): DT, IN, JJ, MD, NN, RB, VB, etc.
Phrase Chunks (C): ADJP, ADVP, NP, PP, VP, etc.
Prefixes (P): *anti, in, mis, non, pre, sub, un,* etc.

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

Table 1: Pivot Templates: Subscript i denotes the index of the head aspect, HA (italicized). Subscript j denotes the index relative to i

Suffixes (S): *able, est, ful, ic, ing, ive, ness* etc.
Word Sentiment Polarity (SP): 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 1). Each template generates several features that capture various positional context around the HA. Additionally, we consider up to 3^{rd} order pivot features allowing us to model various dependencies as features. For polarity, we used the opinion lexicon[2] derived from (Hu and Liu, 2004).

Feature Templates: Table 1 details the templates for features pivoting on the head aspect. Various features from these templates coupled with the value of the current sequence tag at y_t or a combination of current and previous labels y_t, y_{t-1} serve as our linear chain features (LCF), $f(y_{i-1}, y_i, w)$. Further, the index i for LCF can refer to any word in the sentence and not necessarily the head aspect, yielding us a very rich feature space.

Learning the CRF λs: Given a set of training examples $\{\mathbf{w_i}, \overline{y_t}\}$ where $\overline{y_t}$ are the correct sequence tags, we estimate the CRF $\Lambda = \{\lambda_k\}$parameters by minimizing the negative log-likelihood (NLL),

$$\Lambda = argmin_\Lambda(C \sum_i \log(p(\overline{y_t}|\mathbf{w_i}, \Lambda)) + \sum_k \lambda_k^2) \quad (11)$$

Where $p(\overline{y_t}|\mathbf{w}, \Lambda) \propto exp(\sum_k \sum_t \lambda_k f_k(y_{t-1}, y_t, w))$, C is the soft-margin parameter, and the term $\sum_k \lambda_k^2$ indicates L_2 regularization on the feature weights, λ_k. The training set for learning the embedded CRF models λs is detailed in Table 3 (col 1, 2).

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

4 Optimizing Phrasal Collocations

Topic models can be described in terms of a simple Pólya urn (SPU) sampling schemes in the sense that when a particular term (word or phrase) is drawn from a topic, count of that term is incremented in that topic. This enforces the topic distribution to tend towards these terms over time as frequency of them increases. Therefore, the posterior of generative topic models often favors terms with high frequency e.g., unigrams, while phrasal terms are ranked lower due to their lower frequencies. This is undesirable for phrase extraction.

In contrast, Generalized Pőlya urn (GPU) model differs from SPU in its sampling process. When a certain term is drawn, the count of that term increases as well as it also increases the count of terms which are similar to drawn word/phrases via pseudo-counts for promotion. Thus, GPU caters for promotion of others terms in a principled manner. It has been previously used for unigram topic modeling in (Mimno et al., 2011). In this work we leverage it for phrases.

4.1 Proposed PSM-GPU model

We optimize the collocations of relevant aspect words and phrases in the GPU framework in two ways:

Word to phrase: Intuitively, if an aspect word is assigned to a topic then that topic should represent that aspect and to all other phrases in that aspect's phrase set (i.e., phrases containing that aspect) should belong to the same topic. Thus, when an aspect word is assigned to a topic then each phrase in its aspect set is promoted with a small count in that topic.
Phrase to word: When a phrase $\langle w_{d,s,j} \rangle_{j=st}^{j=st+len-1}$ is assigned to a topic, each component word $w_{d,s,j}$ where $j \in [st, st+len-1]$ in it is also promoted with a certain small count, i.e., each word of that phrase is also assigned to that topic by a certain amount.

We now define the term promotion matrix, A for the GPU framework. Every element of $A, A_{w,w'}$ refers to the promotion pseudocount, i.e., whenever a w was seen in an urn, we increment the count by $A_{w,w'}$ of $w^{'}$. $w, w^{'}$ can be word or phrases.

$$A_{w,w'} = \begin{cases} 1 & \text{if } w = w^{'} \\ \sigma & \text{if } w \text{ is an aspect word,} \\ & w^{'} \text{is a phrase} \in \text{Phrase set of } w \\ \delta & w^{'} \text{is a word} \in \text{Phrase } w \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

To improve the ranking of phrases, the value of σ is kept greater than δ. Empirically values are given section 5.1.

4.2 PSM-GPU Inference

Accounting the GPU process above, the approximate Gibbs samplers for z and r take the following form:

$$p(z_{d,s} = a | Z_{\neg d,s}, R_{\neg d,s}, W_{\neg d,s}) \propto \left[\frac{(n^s_{d,a})_{\neg d,s} + \alpha)}{(n^s_{d,(\cdot)})_{\neg d,s} + A\alpha)} \right] \times \frac{\left(\prod_{v=1}^{V} \frac{\Gamma(\sum_{w'=1}^{V} (A_{v,w'} * n^A_{a,w'}) + \beta)}{\Gamma(\sum_{w'=1}^{V} (A_{v,w'} * n^A_{a,w'})_{\neg d,s} + \beta)} \right)}{\left(\frac{\Gamma(\sum_{v=1}^{V} \sum_{w'=1}^{V} (A_{v,w'} * n^A_{a,w'}) + V\beta)}{\Gamma(\sum_{v=1}^{V} \sum_{w'=1}^{V} (A_{v,w'} * n^A_{a,w'})_{\neg d,s} + V\beta)} \right)} \times \left[\left(\prod_{v=1}^{V} \frac{\Gamma(n^B_{a,v} + \beta)}{\Gamma((n^B_{a,v})_{\neg d,s} + \beta)} \right) / \left(\frac{\Gamma(n^B_{a,(\cdot)} + V\beta)}{\Gamma((n^B_{a,(\cdot)})_{\neg d,s} + V\beta)} \right) \right] \quad (13)$$

$$p(r_{d,s,j} = 1 | z_{d,s} = a, w_{d,s,j} = v, \ldots) \propto \frac{\sum_{w'=1}^{V} (A_{v,w'} * n^A_{a,w'})_{\neg d,s,j} + \beta)}{\sum_{v=1}^{V} \sum_{w'=1}^{V} (A_{v,w'} * n^A_{a,w'})_{\neg d,s,j} + V\beta} \times p_{CRF}(r_{d,s,j-1}, r_{d,s,j} = 1 | v) \quad (14)$$

Similarly, phrasal opinion switch variable can be derived. The sampler for individual background words remains unchanged.

5 Experimental Evaluation

In this section, we evaluate our proposed models. We first detail our dataset, followed by baselines and results.

5.1 Dataset and Parameter Settings

Domain	Pos. Labeled Phrase	Neg. Labeled Phrase	Positive	Negative	Total
Router	414	1256	1937	5291	7228
GPS	948	672	2473	2231	4704
Mouse	376	477	1421	2591	4012
Keyboard	398	660	912	1539	2451

Table 3: Statistics of dataset of four domain

Dataset Statistics: For CRF training, we created a phrase labeled dataset of aspect opinion phrases using product reviews from Amazon across 4 domains each spanning 4 head aspects. In this work, head aspects for a domain are known a priori either directly using unsupervised topic induction or guided by domain knowledge (e.g. using aspect models such as (Zhao et al., 2010; Chen et al., 2013; Mukherjee and Liu, 2012)). Our focus is on phrase extraction and grouping. We labeled the positive and negative opinion phrases spans (Table 3; col 2, 3) in the reviews following the annotation schemes in (Wilson et al., 2005) for embedded CRF training in PSM. Table 3 details our labeled data for CRF training. This phrase boundary labeled dataset (Table 3; col 2, 3) is "orthogonal" or disjoint from the data where the PSM model was fit and evaluated (Table 3, col 4, 5). This avoids overfitting and makes a fair case for all the experiments of PSM.
Preprocessing and Parameter Setting: We removed the stopwords, punctuation, special characters and words appearing less than 5 times in each domain. For all models, posterior estimates of latent variables were taken with a sampling lag of 50 iterations post burn-in phase (of 200 iterations) with 2,000 iterations in total. Dirichlet priors were set to $\alpha = 50/K$, where K is the number of topics (empirically set to 10 via pilot) and $\beta = 0.1$. The CRF parameters $C = 1$ and GPU parameters $\sigma = 0.05$ and $\delta = 0.01$ were estimated using cross validation.

5.2 Baselines

We consider the following relevant phrase extraction models as our baselines:

PSM-GPU	sMC-GPU	LDA-P-GPU
Router→Connection:"updating firmware secure connection", "dropping connection", "connection excellent", "instability wireless connection", "crashes entire connection", *"updating"*, "kills current connection including downloads", "affected plugged connection",*"cable radio frequency connection", "halfway"*	**Router→Connection:** *"connection big time", "signal weak time connection", "connection time stable"*, "lose connection", "internet connection speed dropped", *"stop working lot connection"*, "started dropping internet connection", "started dropping connection", "drop wireless connection drops wired connection", "connection dropping problems"	**Router→Connection:** *"drops", "times day internet connection"*, "internet connection multiple times", *"internet connection times", "internet connection minutes"*, "dropping internet connection", "broadband internet connection", "extremely slow internet connection", "lost internet connection",*"internet connection couple"*
GPS→Screen:"poor screen contrast daylight", *"direction screen missing poor screen contrast"*, "slow screen size makes useless", "screen turned", *"turn", "nothing screen"*, "night screen bit bright", "screen unpredictable directions", "lacks faster screen refresh rate", "smoother screen refresh"	**GPS→Screen:** *"bright"*, "excellent nice touch screen", *"touch screen n't"*, "nice big touch screen", "touch screen big size", *"smaller", "touch screen but nice", "screen accurate spoken direction"*, "touch screen nice", "nice slim touch screen"	**GPS→Screen:** *"smaller screen but dont"*, "wide screen", *"nothing but screen", "ok but screen", "screen but normal", "screen size"*, "large screen", "better traffic features cons touch screen", *"screen real estate than", "than years screen"*

Table 2: Example aspect specific opinion phrases (comma delimited in order) discovered by PSM-GPU, sMC-GPU, LDA-P-GPU. Errors are italicized and marked in red.

LDA with phrases (LDA-P): As aspect-sentiment phrases are often noun phrases, a basic approach is to include the noun phrases (extracted using a parser) as separate terms in the corpus.

Topical N-gram (TNG): The TNG model in (Wang et al., 2007) extends LDA to model n-grams of arbitrary length. As aspects often appear close to their opinion in the sentence, topical n-grams for each aspect form a natural baseline. We used the authors original implementation in the MALLET toolbox.

LDA-P with GPU (LDA-P-GPU): This model is due to (Fei et al., 2014) and is tailored for phrase extraction in opinion mining. It employs LDA with noun phrases in the GPU framework to rank the aspect phrases higher in their topics. Our implementations of LDA-P and LDA-P-GPU use the noun phrases discovered by the Stanford Parser.

semi-Markov CRF with GPU(sMC-GPU): This model builds over the model of (Yang and Cardie, 2012) that used dependency tree features and semi-CRF to model the arbitrarily long expressions. We used these expression spans as multiword in vocabulary. Then we employ GPU based sampling with LDA proposed by (Fei et al., 2014) to collocate opinion expressions.

5.3 Qualitative Analysis

To assess the quality of extracted expressions, we labeled the topics following instructions in (Mimno et al., 2011). First, each topic was labeled as coherent or incoherent and an aspect name was given if the topic was coherent. Each topic was presented as a list of top 45 terms in descending order of their probabilities under that topic. A topic was considered coherent if the terms in the topic were semantically related to each other.

Next, for coherent topics, their terms were labeled as correct (if the terms semantics was relevant to the topic) or incorrect (otherwise). Two human judges were used in the annotation. Agreements being high ($\kappa > 0.78$), disagreements were resolved upon consensus among judges.

Table 2 reports the top 10 terms(words/phrases) for aspect 'connection' (Router domain) and aspect 'screen'(GPS domain) across PSM-GPU, sMC-GPU and LDA-P-GPU (the two closest competitor). We note that PSM-GPUs phrases are more expressive compared to sMC-GPU because sMC-GPU is prone to have longer phrases due to segment features but PSM's switch variable captured more relevant aspect specific opinion expressions. sMC-GPU has better

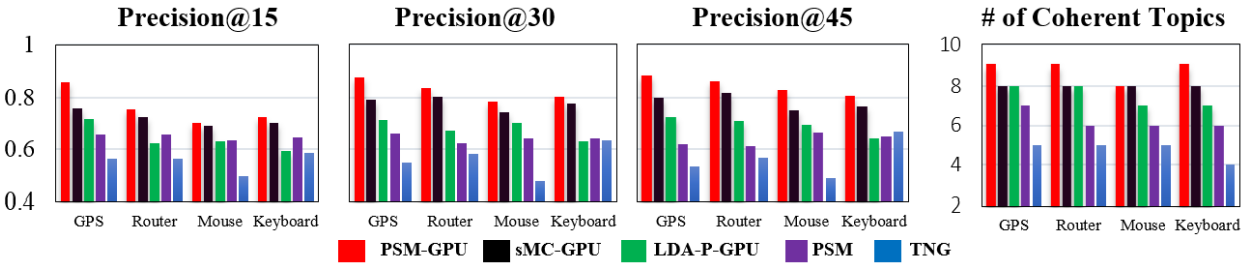

Figure 2: Charts from left to right are Topical words Precision@15, Precision@30, Precision@45 of coherent topics of each model and last one is number of coherent topics of each model.

Domain	PSM-GPU				PSM				sMC-GPU				LDA-P-GPU				TNG				LDA-P			
	P	R	F	Ac.	P	R	F	Ac.	P	R	F	Ac.	P	R	F	Ac.	P	R	F	Ac.	P	R	F	Ac.
Router	71.8	67.8	69.7	68.8	69.0	67.3	68.1	67.7	69.6	67.7	68.6	68.2	65.2	67.9	66.5	67.1	65.2	66.8	66.0	65.9	65.5	68.0	66.7	67.4
GPS	69.3	69.2	69.2	69.3	66.7	67.7	67.2	67.4	65.3	68.0	65.7	65.5	63.6	68.4	65.9	66.8	64.1	68.3	66.1	58.8	66.0	69.7	67.8	68.6
Mouse	87.4	81.4	84.0	87.3	82.9	80.4	81.3	82.4	83.4	81.5	82.6	82.3	81.3	78.3	79.4	80.9	82.9	77.9	80.0	82.5	77.9	77.3	77.1	77.7
Keyboard	92.5	72.9	81.2	84.1	90.2	70.1	78.4	81.6	88.4	70.3	77.6	81.0	86.7	69.3	76.4	79.9	88.8	66.6	75.8	79.1	83.6	63.4	71.4	75.0
Avg.	80.3	72.8	**76.0**	77.4	77.2	71.4	**73.8**	74.8	76.7	71.9	**73.6**	74.3	74.2	71.0	**72.1**	73.7	75.3	69.9	**72.0**	71.6	73.3	69.6	**70.8**	72.2

Table 4: Sentiment classification: Precision, Recall, F1 and accuracy from top to down for each domain and each model

phrases compared to LDA-P-GPU because the latter only considers noun phrases which may not always be semantically coherent under an aspect. The qualitative results of other baselines TNG and LDA-P were worse than that of LDA-P-GPU and hence omitted due to space constraints. However, the subsequent experiments compare all models quantitatively.

5.4 Quantitative Analysis

We consider the following metrics and tasks:

Average Precision: Figure 2 shows the average Precision@n (p@n) for $n = 15, 30, 45$ of all coherent topics for each model in each domain. We note that PSM-GPU achieves the highest precision for all domains significantly ($p < 0.01$) outperforming its closest competitor sMC-GPU. sMC-GPU tends to discover longer phrases due to segment features in semi-CRF and combined with GPU gains the maximum strength among other baselines. Next in order are LDA-P-GPU, PSM, and TNG. We have not shown the result of LDA-P as its top terms didn't contain enough phrases and its precision scores were quite lower compared to other models. But it is worthwhile to note that PSM outperforms LDA-P-GPU (2nd best competitor) at lower ranks which is more important (e.g., in majority domains for $p@15$) and shows its effectiveness. It is a bit unfair to compare PSM with sMC-GPU because PSM is lacking phrase rank optimization whereas sMC-GPU enforces it, and the $p@n$ metric uses rank position as its goodness criterion. However, we will see that in an actual application task, both PSM and PSM-GPU does better than sMC-GPU. Also, we observed that p@45 is higher compare to p@15 or p@30. The reason is even though we are promoting the phrases using GPU it is not able to remove some aspect opinion words from top 15 terms due to their high occurrence in phrases. For e.g. Table 2 has opinion words like "updating", "turn" which are considered incorrect because of non-phrasal terms.

# of coherent topics: Figure 2 (rightmost chart) shows the number of coherent topics produced by each model. A model that can discover more coherent topics is better. We find that PSM-GPU can discover more coherent topics with phrases than its baselines across all domains. The trends of other models are similar to $p@n$ and can be analogously explained.

We note that the Topic Coherence (TC) metric in (Mimno et al., 2011) which is often used to approximate coherence in unigram topic models as it correlates with human notions of coherence, uses co-document frequency of individual words in topics.

However, in our problem as phrases are sparse, their co-document frequency is far lower than words. Hence, the TC metric is not directly applicable. Our measure of coherence is based on human judgment (achieves high agreements, $\kappa > 0.78$ see Section 5.3) and from Table 2 we can see the discovered phrases do reflect coherence. Hence, to evaluate the phrases quantitatively, we employ an actual sentiment classification task that uses the posterior of our models (top phrases) as features. This is reasonable because the estimated topics (when used as features) improve sentiment classification, it shows that they are meaningful and capable of capturing latent sentiment that govern polarities.

Sentiment Classification: For this task, instead of using all the words as features, we used the posterior on φ^A (top 50 terms of φ^A) as features. For all models, all possible n-grams of top 50 terms are also considered as features. We trained SVMs[3] (using the SVMLight toolkit) with the features described above. Evaluation for this task employed 5-fold cross validation on the data in Table 3 (col 4, 5). For each test fold, the features were induced upon fitting the aspect extraction models on the training data of that fold. From Table 4 we note that both PSM-GPU and PSM outperform all competitors on average F1 across all domains. More specifically, we note that PSM alone that uses no rank optimization performs better than sMC-GPU employing phrasal rank optimization under GPU scheme. We believe this is due to PSM's switching component that can discover correct aspect/sentiment terms (sufficient for polarity classification) and rank it higher based on frequency even though the expressive aspect specific phrases remain ranked lower. sMC-GPU tends to have longer phrases so it does well, however, under GPU, longer phrases may not be promoted well as they lack anchor aspect terms under a relevant topic. LDA-P-GPU uses standard (Noun Phrases) NPs for phrases with rank optimization and hence is the next in performance order as NPs may not capture opinion well. TNG does not perform as well as it relies on multiword collocation as opposed to NP/VP for phrase extraction. LDA-P's performance is lowest as it cannot rank the relevant NPs high. PSM-GPU has the right balance of phrase boundary span and phrasal rank optimization via GPU that makes it significantly outperform ($p < 0.01$) all competitors.

[3]Using an RBF kernel ($C = 10, g = 0.01$) which performed best upon tuning various SVM parameters via cross validation.

Domain	Precision	Recall	F1	Acc.
Router	69.68	67.3	68.4	67.9
GPS	66.47	68.6	67.5	68.00
Mouse	87.24	80.5	83.4	86.9
Keyboard	85.45	70.9	77.1	81.1

Table 5: Domain ablation result on polarity classification

Sequence Model Sensitivity: To assess the robustness of the hybrid framework, we evaluate the sensitivity of the embedded CRF model via domain ablation. We choose the best performer PSM-GPU and ablate each domain during its CRF training. We repeat the previous experiment on sentiment classification using the ablated model. From the results in Table 5, we note that the reduction in precision is relatively more than that of recall. However, the F1 score does not drop significantly (compared to Table 4) for any domain showing the robustness of the hybrid framework. We note that even with some skewness in the labeled data (Table 3), CRF is not overfitting here and the proposed pivot features (Table 1) are powerful enough to learn the phrasal structure across domain.

6 Conclusion

This paper presented a novel hybrid framework for aspect specific opinion expression extraction. Two models PSM and PSM-GPU were proposed that employ CRF discriminative sequence modeling for phrase boundary extraction and generative modeling for grouping relevant terms under a topic. PSM-GPU further optimized the aspect coherence using the generalized Pólya urn sampling scheme. Experimental results showed that the proposed hybrid models can extract more coherent aspect specific opinion expressions significantly outperforming all competitors across all domains and are robust in cross-domain knowledge transfer.

Acknowledgment

This work was supported in part by NSF 1527364.

References

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *IJCNLP*, pages 1162–1170. Citeseer.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *IJCAI*, volume 7, pages 2683–2688.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Exploiting domain knowledge in aspect extraction. In *EMNLP*, pages 1655–1667.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439. Association for Computational Linguistics.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316.

Geli Fei, Zhiyuan Chen, and Bing Liu. 2014. Review topic discovery with phrases using the pólya urn model. In *COLING*, pages 667–676.

Geli Fei, Zhiyuan Brett Chen, Arjun Mukherjee, and Bing Liu. 2016. Discovering correspondence of sentiment words and aspects. In *In proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1035–1045. Association for Computational Linguistics.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Richard Johansson and Alessandro Moschitti. 2011. Extracting opinion expressions and their polarities: exploration of pipelines and joint models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 101–106. Association for Computational Linguistics.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP-CoNLL*, volume 7, pages 1065–1074. Citeseer.

T Kudo. 2009. Crf++: Yet another crf toolkit [ol].

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Huayi Li, Arjun Mukherjee, Jianfeng Si, and Bing Liu. 2015. Extracting verb expressions implying negative opinions. In *AAAI*, pages 2411–2417.

Robert V Lindsey, William P Headden III, and Michael J Stipicevic. 2012. A phrase-discovering topic model using hierarchical pitman-yor processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 214–222. Association for Computational Linguistics.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744. ACM.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 339–348. Association for Computational Linguistics.

Arjun Mukherjee. 2016. Extracting aspect specific sentiment expres-sions implying negative opinions. In *In proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics*.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, pages 486–495.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2011. Content models with attitude. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 350–358. Association for Computational Linguistics.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 697–702. IEEE.

Shuai Wang, Zhiyuan Chen, and Bing Liu. 2016. Mining aspect-specific opinion using a holistic lifelong topic model. In *Proceedings of the 25th International Conference on World Wide Web*, pages 167–176. International World Wide Web Conferences Steering Committee.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1533–1541. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *ACL (1)*, pages 1640–1649.

Bishan Yang and Claire Cardie. 2014. Joint modeling of opinion expression extraction and attribute classification. *Transactions of the Association for Computational Linguistics*, 2:505–516.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65. Association for Computational Linguistics.

Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. 2011. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 379–388. Association for Computational Linguistics.