Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization

Bishan Yang
Department of Computer Science
Cornell University
bishan@cs.cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
cardie@cs.cornell.edu

Abstract

This paper proposes a novel context-aware method for analyzing sentiment at the level of individual sentences. Most existing machine learning approaches suffer from limitations in the modeling of complex linguistic structures across sentences and often fail to capture non-local contextual cues that are important for sentiment interpretation. In contrast, our approach allows structured modeling of sentiment while taking into account both local and global contextual information. Specifically, we encode intuitive lexical and discourse knowledge as expressive constraints and integrate them into the learning of conditional random field models via posterior regularization. The context-aware constraints provide additional power to the CRF model and can guide semi-supervised learning when labeled data is limited. Experiments on standard product review datasets show that our method outperforms the state-of-the-art methods in both the supervised and semi-supervised settings.

1 Introduction

The ability to extract sentiment from text is crucial for many opinion-mining applications such as opinion summarization, opinion question answering and opinion retrieval. Accordingly, extracting sentiment at the fine-grained level (e.g. at the sentence- or phrase-level) has received increasing attention recently due to its challenging nature and its importance in supporting these opinion analysis tasks (Pang and Lee, 2008).

In this paper, we focus on the task of sentence-level sentiment classification in online reviews. Typical approaches to the task employ supervised machine learning algorithms with rich features and take into account the interactions between words to handle compositional effects such as polarity reversal (e.g. (Nakagawa et al., 2010; Socher et al., 2013)). Still, their methods can encounter difficulty when the sentence on its own does not contain strong enough sentiment signals (due to the lack of statistical evidence or the requirement for background knowledge). Consider the following review for example,

> 1. Hearing the music in real stereo is a true revelation. 2. You can feel that the music is no longer constrained by the mono recording. 3. In fact, it is more like the players are performing on a stage in front of you ...

Existing feature-based classifiers may be effective in identifying the positive sentiment of the first sentence due to the use of the word *revelation*, but they could be less effective in the last two sentences due to the lack of explicit sentiment signals. However, if we examine these sentences within the discourse context, we can see that: the second sentence expresses sentiment towards the same aspect – *the music* – as the first sentence; the third sentence expands the second sentence with the discourse connective *In fact*. These discourse-level relations help indicate that sentence 2 and 3 are likely to have positive sentiment as well.

The importance of discourse for sentiment analysis has become increasingly recognized. Most existing work considers discourse relations between adjacent sentences or clauses and incorporates them as constraints (Kanayama and Nasukawa, 2006; Zhou et al., 2011) or features in classifiers Trivedi and Eisenstein (2013; Lazaridou et al. (2013). Very little work has explored long-distance discourse relations for sentiment analysis. Somasundaran et al. (2008) defines coreference relations on opinion targets and applies them to constrain the polarity of sentences.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 325–335,
Baltimore, Maryland, USA, June 23-25 2014.

However, the discourse relations were obtained from fine-grained annotations and implemented as hard constraints on polarity.

Obtaining sentiment labels at the fine-grained level is costly. Semi-supervised techniques have been proposed for sentence-level sentiment classification (Täckström and McDonald, 2011a; Qu et al., 2012). However, they rely on a large amount of document-level sentiment labels that may not be naturally available in many domains.

In this paper, we propose a sentence-level sentiment classification method that can (1) incorporate rich discourse information at both local and global levels; (2) encode discourse knowledge as soft constraints during learning; (3) make use of unlabeled data to enhance learning. Specifically, we use the Conditional Random Field (CRF) model as the learner for sentence-level sentiment classification, and incorporate rich discourse and lexical knowledge as soft constraints into the learning of CRF parameters via Posterior Regularization (PR) (Ganchev et al., 2010). As a framework for structured learning with constraints, PR has been successfully applied to many structural NLP tasks (Ganchev et al., 2009; Ganchev et al., 2010; Ganchev and Das, 2013). Our work is the first to explore PR for sentiment analysis. Unlike most previous work, we explore a rich set of structural constraints that cannot be naturally encoded in the feature-label form, and show that such constraints can improve the performance of the CRF model.

We evaluate our approach on the sentence-level sentiment classification task using two standard product review datasets. Experimental results show that our model outperforms state-of-the-art methods in both the supervised and semi-supervised settings. We also show that discourse knowledge is highly useful for improving sentence-level sentiment classification.

2 Related Work

There has been a large amount of work on sentiment analysis at various levels of granularity (Pang and Lee, 2008). In this paper, we focus on the study of sentence-level sentiment classification. Existing machine learning approaches for the task can be classified based on the use of two ideas. The first idea is to exploit sentiment signals at the sentence level by learning the relevance of sentiment and words while taking into account the context in which they occur: Nakagawa et al. (2010) uses tree-CRF to model word interactions based on dependency tree structures; Choi and Cardie (2008) applies compositional inference rules to handle polarity reversal; Socher et al. (2011) and Socher et al. (2013) compute compositional vector representations for words and phrases and use them as features in a classifier.

The second idea is to exploit sentiment signals at the inter-sentential level. Polanyi and Zaenen (2006) argue that discourse structure is important in polarity classification. Various attempts have been made to incorporate discourse relations into sentiment analysis: Pang and Lee (2004) explored the consistency of subjectivity between neighboring sentences; Mao and Lebanon (2007),McDonald et al. (2007), and Täckström and McDonald (2011a) developed structured learning models to capture sentiment dependencies between adjacent sentences; Kanayama and Nasukawa (2006) and Zhou et al. (2011) use discourse relations to constrain two text segments to have either the same polarity or opposite polarities; Trivedi and Eisenstein (2013) and Lazaridou et al. (2013) encode the discourse connectors as model features in supervised classifiers. Very little work has explored long-distance discourse relations. Somasundaran et al. (2008) define opinion target relations and apply them to constrain the polarity of text segments annotated with target relations. Recently, Zhang et al. (2013) explored the use of explanatory discourse relations as soft constraints in a Markov Logic Network framework for extracting subjective text segments.

Leveraging both ideas, our approach exploits sentiment signals from both intra-sentential and inter-sentential context. It has the advantages of utilizing rich discourse knowledge at different levels of context and encoding it as soft constraints during learning.

Our approach is also semi-supervised. Compared to the existing work on semi-supervised learning for sentence-level sentiment classification (Täckström and McDonald, 2011a; Täckström and McDonald, 2011b; Qu et al., 2012), our work does not rely on a large amount of coarse-grained (document-level) labeled data, instead, distant supervision mainly comes from linguistically-motivated constraints.

Our work also relates to the study of posterior regularization (PR) (Ganchev et al., 2010). PR has been successfully applied to many structured NLP

tasks such as dependency parsing, information extraction and cross-lingual learning tasks (Ganchev et al., 2009; Bellare et al., 2009; Ganchev et al., 2010; Ganchev and Das, 2013). Most previous work using PR mainly experiments with feature-label constraints. In contrast, we explore a rich set of linguistically-motivated constraints which cannot be naturally formulated in the feature-label form. We also show that constraints derived from the discourse context can be highly useful for disambiguating sentence-level sentiment.

3 Approach

In this section, we present the details of our proposed approach. We formulate the sentence-level sentiment classification task as a sequence labeling problem. The inputs to the model are sentence-segmented documents annotated with sentence-level sentiment labels (positive, negative or neutral) along with a set of unlabeled documents. During prediction, the model outputs sentiment labels for a sequence of sentences in the test document. We utilize conditional random fields and use Posterior Regularization (PR) to learn their parameters with a rich set of context-aware constraints.

In what follows, we first briefly describe the framework of Posterior Regularization. Then we introduce the context-aware constraints derived based on intuitive discourse and lexical knowledge. Finally we describe how to perform learning and inference with these constraints.

3.1 Posterior Regularization

PR is a framework for structured learning with constraints (Ganchev et al., 2010). In this work, we apply PR in the context of CRFs for sentence-level sentiment classification.

Denote $\mathbf{x}$ as a sequence of sentences within a document and $\mathbf{y}$ as a vector of sentiment labels associated with $\mathbf{x}$. The CRF model the following conditional probabilities:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{\exp(\theta \cdot f(\mathbf{x}, \mathbf{y}))}{Z_\theta(\mathbf{x})}$$

where $f(\mathbf{x}, \mathbf{y})$ are the model features, θ are the model parameters, and $Z_\theta(\mathbf{x}) = \sum_{\mathbf{y}} \exp(\theta \cdot f(\mathbf{x}, \mathbf{y}))$ is a normalization constant. The objective function for a standard CRF is to maximize the log-likelihood over a collection of labeled documents plus a regularization term:

$$\max_\theta \mathcal{L}(\theta) = \max_\theta \sum_{(\mathbf{x},\mathbf{y})} \log p_\theta(\mathbf{y}|\mathbf{x}) - \frac{||\theta||_2^2}{2\delta^2}$$

PR makes the assumption that the labeled data we have is not enough for learning good model parameters, but we have a set of constraints on the posterior distribution of the labels. We can define the set of desirable posterior distrbutions as

$$\mathcal{Q} = \{q(\mathbf{Y}) : E_q[\phi(\mathbf{X}, \mathbf{Y})] = \mathbf{b}\} \quad (1)$$

where ϕ is a constraint function, $\mathbf{b}$ is a vector of desired values of the expectations of the constraint functions under the distribution q [1]. Note that the distribution q is defined over a collection of unlabeled documents where the constraint functions apply, and we assume independence between documents.

The PR objective can be written as the original model objective penalized with a regularization term, which minimizes the KL-divergence between the desired model posteriors and the learned model posteriors with an L2 penalty [2] for the constraint violations.

$$\begin{aligned}\max_\theta \mathcal{L}(\theta) - \min_{q\in\mathcal{Q}} \{KL(q(\mathbf{Y})||p_\theta(\mathbf{Y}|\mathbf{X})) \\ + \beta||E_q[\phi(\mathbf{X}, \mathbf{Y})] - \mathbf{b}||_2^2\}\end{aligned} \quad (2)$$

The objective can be optimized by an EM-like scheme that iteratively solves the minimization problem and the maximization problem. Solving the minimization problem is equivalent to solving its dual since the objective is convex. The dual problem is

$$\arg\max_\lambda \lambda \cdot \mathbf{b} - \log Z_\lambda(X) - \frac{1}{4\beta}||\lambda||_2^2 \quad (3)$$

We optimize the objective function 2 using stochastic projected gradient, and compute the learning rate using AdaGrad (Duchi et al., 2010).

3.2 Context-aware Posterior Constraints

We develop a rich set of context-aware posterior constraints for sentence-level sentiment analysis by exploiting lexical and discourse knowledge. Specifically, we construct the lexical constraints by extracting sentiment-bearing patterns

[1] In general, inequality constraints can also be used. We focus on the equality constraints since we found them to express the sentiment-relevant constraints well.

[2] Other convex functions can be used for the penalty. We use L2 norm because it works well in practice. β is a regularization constant

within sentences and construct the discourse-level constraints by extracting discourse relations that indicate sentiment coherence or sentiment changes both within and across sentences. Each constraint can be formulated as equality between the expectation of a constraint function value and a desired value set by prior knowledge. The equality is not strictly enforced (due to the regularization in the PR objective 2). Therefore all the constraints are applied as soft constraints. Table 1 provides intuitive description and examples for all the constraints used in our model.

Lexical Patterns The existence of a polarity-carrying word alone may not correctly indicate the polarity of the sentence, as the polarity can be reversed by other polarity-reversing words. We extract lexical patterns that consist of polar words and negators [3], and apply the heuristics based on compositional semantics (Choi and Cardie, 2008) to assign a sentiment value to each pattern.

We encode the extracted lexical patterns along with their sentiment values as feature-label constraints. The constraint function can be written as

$$\phi_w(x, y) = \sum_i f_w(x_i, y_i)$$

where $f_w(x_i, y_i)$ is a feature function which has value 1 when sentence x_i contains the lexical pattern w and its sentiment label y_i equals to the expected sentiment value and has value 0 otherwise. The constraint expectation value is set to be the prior probability of associating w with its sentiment value. Note that sentences with neutral sentiment can also contain such lexical patterns. Therefore we allow the lexical patterns to be assigned a neutral sentiment with a prior probability r_0 (we compute this value as the empirical probability of neutral sentiment in the training documents). Using the polarity indicated by lexical patterns to constrain the sentiment of sentences is quite aggressive. Therefore we only consider lexical patterns that are strongly discriminative (many opinion words in the lexicon only indicate sentiment with weak strength). The selected lexical patterns include a handful of seed patterns (such as "pros" and "cons") and the lexical patterns that have high precision (larger then 0.9) of predicting sentiment in the training data.

[3]The polar words are identified using the MPQA lexicon and the negators are identified using a handful of seed words extended by the General Inquirer dictionary and WordNet as described in (Choi and Cardie, 2008).

Discourse Connectives. Lexical patterns can be limited in capturing contextual information since they only look at interactions between words within an expression. To capture context at the clause or sentence level, we consider discourse connectives, which are cue phrases or words that indicate discourse relations between adjacent sentences or clauses. To identify discourse connectives, we apply a discourse tagger trained on the Penn Discourse Treebank (Prasad et al., 2008) [4] to our data. Discourse connectives are tagged with four senses: *Expansion, Contingency, Comparison, Temporal*.

Discourse connectives can operate at both intra-sentential and inter-sentential level. For example, the word "although" is often used to connect two polar clauses within a sentence, while the word "however" is often used to at the beginning of the sentence to connect two polar sentences. It is important to distinguish these two types of discourse connectives. We consider a discourse connective to be intra-sentential if it has the *Comparison* sense and connects two polar clauses with opposite polarities (determined by the lexical patterns). We construct a feature-label constraint for each intra-sentential discourse connective and set its expected sentiment value to be neutral.

Unlike the intra-sentential discourse connectives, the inter-sentential discourse connectives can indicate sentiment transitions between sentences. Intuitively, discourse connectives with the senses of *Expansion* (e.g. also, for example, furthermore) and *Contingency* (e.g. as a result, hence, because) are likely to indicate sentiment coherence; discourse connectives with the sense of *Comparison* (e.g. but, however, nevertheless) are likely to indicate sentiment changes. This intuition is reasonable but it assumes the two sentences connected by the discourse connective are both polar sentences. In general, discourse connectives can also be used to connect non-polar (neutral) sentences. Thus it is hard to directly constrain the posterior expectation for each type of sentiment transitions using inter-sentential discourse connectives.

Instead, we impose constraints on the model posteriors by reducing constraint violations. We

[4]http://www.cis.upenn.edu/~epitler/discourse.html

Types	Description and Examples	Inter-sentential
Lexical patterns	The sentence containing a polar lexical pattern w tends to have the polarity indicated by w. Example lexical patterns are *annoying, hate, amazing, not disappointed, no concerns, favorite, recommend*.	
Discourse Connectives (clause)	The sentence containing a discourse connective c which connects its two clauses that have opposite polarities indicated by the lexical patterns tends to have neutral sentiment. Example connectives are *while, although, though, but*.	
Discourse Connectives (sentence)	Two adjacent sentences which are connected by a discourse connective c tends to have the same polarity if c indicates a *Expansion* or *Contingency* relation, e.g. *also, for example, in fact, because* ; opposite polarities if c indicates a *Comparison* relation, e.g. *otherwise, nevertheless, however*.	✓
Coreference	The sentences which contain coreferential entities appeared as targets of opinion expressions tend to have the same polarity.	✓
Listing patterns	A series of sentences connected via a listing tend to have the same polarity.	✓
Global labels	The sentence-level polarity tends to be consistent with the document-level polarity.	✓

Table 1: Summarization of Posterior Constraints for Sentence-level Sentiment Classification

define the following constraint function:

$$\phi_{c,s}(x, y) = \sum_i f_{c,s}(x_i, y_i, y_{i-1})$$

where c denotes a discourse connective, s indicates its sense, and $f_{c,s}$ is a penalty function that takes value 1.0 when y_i and y_{i-1} form a contradictory sentiment transition, that is, $y_i \neq_{polar} y_{i-1}$ if $s \in \{\textit{Expansion}, \textit{Contingency}\}$, or $y_i =_{polar} y_{i-1}$ if $s = \textit{Comparison}$. The desired value for the constraint expectation is set to 0 so that the model is encouraged to have less constraint violations.

Opinion Coreference Sentences in a discourse can be linked by many types of coherence relations (Jurafsky et al., 2000). Coreference is one of the commonly used relations in written text. In this work, we explore coreference in the context of sentence-level sentiment analysis. We consider a set of polar sentences to be linked by the *opinion coreference* relation if they contain coreferring opinion-related entities. For example, the following sentences express opinions towards "the speaker phone", "The speaker phone" and "it" respectively. As these opinion targets are coreferential (referring to the same entity "the speaker phone"), they are linked by the *opinion coreference* relation [5].

> My favorite features are **the speaker phone** and the radio. **The speaker phone** is very functional. I use **it** in the car, very audible even with freeway noise.

[5]In general, the opinion-related entities include both the opinion targets and the opinion holders. In this work, we only consider the targets since we experiment with single-author product reviews. The opinion holders can be included in a similar way as the opinion targets.

Our coreference relations indicated by opinion targets overlap with the *same target* relation introduced in (Somasundaran et al., 2009). The differences are: (1) we encode the coreference relations as soft constraints during learning instead of applying them as hard constraints during inference time; (2) our constraints can apply to both polar and non-polar sentences; (3) our identification of coreference relations is automatic without any fine-grained annotations for opinion targets.

To extract coreferential opinion targets, we apply Stanford's coreference system (Lee et al., 2013) to extract coreferential mentions in the document, and then apply a set of syntactic rules to identify opinion targets from the extracted mentions. The syntactic rules correspond to the shortest dependency paths between an opinion word and an extracted mention. We consider the 10 most frequent dependency paths in the training data. Example dependency paths include *nsubj*(opinion, mention), *nobj*(opinion, mention), and *amod*(mention, opinion).

For sentences connected by the opinion coreference relation, we expect their sentiment to be consistent. To encode this intuition, we define the following constraint function:

$$\phi_{coref}(x, y) = \sum_{i, ant(i)=j, j \geq 0} f_{coref}(x_i, x_j, y_i, y_j)$$

where $ant(i)$ denotes the index of the sentence which contains an antecedent target of the target mentioned in sentence i (the antecedent relations over pairs of opinion targets can be constructed using the coreference resolver), and f_{coref} is a penalty function which takes value 1.0 when the expected sentiment coherency is violated, that is, $y_i \neq_{polar} y_j$. Similar to the inter-sentential dis-

course connectives, modeling opinion coreference via constraint violations allows the model to handle neutral sentiment. The expected value of the constraint functions is set to 0.

Listing Patterns Another type of coherence relations we observe in online reviews is listing, where a reviewer expresses his/her opinions by listing a series of statements followed by a sequence of numbers. For example, "1. It's smaller than the ipod mini .... 2. It has a removable battery ....". We expect sentences connected by a listing to have consistent sentiment. We implement this constraint in the same form as the coreference constraint (the antecedent assignments are constructed from the numberings).

Global Sentiment Previous studies have demonstrated the value of document-level sentiment in guiding the semi-supervised learning of sentence-level sentiment (Täckström and McDonald, 2011b; Qu et al., 2012). In this work, we also take into account this information and encode it as posterior constraints. Note that these constraints are not necessary for our model and can be applied when the document-level sentiment labels are naturally available.

Based on an analysis of the Amazon review data, we observe that sentence-level sentiment usually doesn't conflict with the document-level sentiment in terms of polarity. For example, the proportion of negative sentences in the positive documents is very small compared to the proportion of positive sentences. To encode this intuition, we define the following constraint function:

$$\phi_g(x, y) = \sum_i^n \delta(y_i \neq_{polar} g)/n$$

where $g \in \{positive, negative\}$ denotes the sentiment value of a polar document, n is the total number of sentences in x, and δ is an indicator function. We hope the expectation of the constraint function takes a small value. In our experiments, we set the expected value to be the empirical estimate of the probability of "conflicting" sentiment in polar documents using the training data.

3.3 Training and Inference

During training, we need to compute the constraint expectations and the feature expectations under the auxiliary distribution q at each gradient step. We can derive q by solving the dual problem in 3:

$$q(\mathbf{y}|\mathbf{x}) = \frac{exp(\theta \cdot f(\mathbf{x}, \mathbf{y}) + \lambda \cdot \phi(\mathbf{x}, \mathbf{y}))}{Z_{\lambda,\theta}(X)} \quad (4)$$

where $Z_{\lambda,\theta}(X)$ is a normalization constant. Most of our constraints can be factorized in the same way as factorizing the model features in the first-order CRF model, and we can compute the expectations under q very efficiently using the forward-backward algorithm. However, some of our discourse constraints (opinion coreference and listing) can break the tractable structure of the model. For constraints with higher-order structures, we use Gibbs Sampling (Geman and Geman, 1984) to approximate the expectations. Given a sequence $\mathbf{x}$, we sample a label $\mathbf{y}_i$ at each position i by computing the unnormalized conditional probabilities $p(\mathbf{y}_i = l|\mathbf{y}_{-i}) \propto exp(\theta \cdot f(\mathbf{x}, \mathbf{y}_i = l, \mathbf{y}_{-i}) + \lambda \cdot \phi(\mathbf{x}, \mathbf{y}_i = l, \mathbf{y}_{-i}))$ and renormalizing them. Since the possible label assignments only differ at position i, we can make the computation efficient by maintaining the structure of the coreference clusters and precomputing the constraint function for different types of violations.

During inference, we find the best label assignment by computing $\arg\max_{\mathbf{y}} q(\mathbf{y}|\mathbf{x})$. For documents where the higher-order constraints apply, we use the same Gibbs sampler as described above to infer the most likely label assignment, otherwise, we use the Viterbi algorithm.

4 Experiments

We experimented with two product review datasets for sentence-level sentiment classification: the Customer Review (CR) data (Hu and Liu, 2004)[6] which contains 638 reviews of 14 products such as cameras and cell phones, and the Multi-domain Amazon (MD) data from the test set of Täckström and McDonald (2011a) which contains 294 reivews from 5 different domains. As in Qu et al. (2012), we chose the books, electronics and music domains for evaluation. Each domain also comes with 33,000 extra reviews with only document-level sentiment labels.

We evaluated our method in two settings: supervised and semi-supervised. In the supervised setting, we treated the test data as unlabeled data and performed transductive learning. In the semi-supervised setting, our unlabeled data consists of

[6]Available at `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`.

both the available unlabeled data and the test data. For each domain in the MD dataset, we made use of no more than 100 unlabeled documents in which our posterior constraints apply. We adopted the evaluation schemes used in previous work: 10-fold cross validation for the CR dataset and 3-fold cross validation for the MD dataset. We also report both two-way classification (positive vs. negative) and three-way classification results (positive, negative or neutral). We use accuracy as the performance measure. In our tables, boldface numbers are statistically significant by paired t-test for $p < 0.05$ against the best baseline developed in this paper [7].

We trained our model using a CRF incorporated with the proposed posterior constraints. For the CRF features, we include the tokens, the part-of-speech tags, the prior polarities of lexical patterns indicated by the opinion lexicon and the negator lexicon, the number of positive and negative tokens and the output of the vote-flip algorithm (Choi and Cardie, 2009). In addition, we include the discourse connectives as local or transition features and the document-level sentiment labels as features (only available in the MD dataset).

We set the CRF regularization parameter $\sigma = 1$ and set the posterior regularization parameter β and γ (a trade-off parameter we introduce to balance the supervised objective and the posterior regularizer in 2) by using grid search [8]. For approximation inference with higher-order constraints, we perform 2000 Gibbs sampling iterations where the first 1000 iterations are burn-in iterations. To make the results more stable, we construct three Markov chains that run in parallel, and select the sample with the largest objective value.

All posterior constraints were developed using the training data on each training fold. For the MD dataset, we also used the dvd domain as additional labeled data for developing the constraints.

Baselines. We compared our method to a number of baselines: (1) CRF: CRF with the same set of model features as in our method. (2) CRF-INF: CRF augmented with inference constraints. We can incorporate the proposed constraints (constraints derived from lexical patterns and discourse connectives) as hard constraints into CRF during inference by manually setting λ in equation 4 to a large value,[9]. When λ is large enough, it is equivalent to adding hard constraints to the viterbi inference. To better understand the different effects of lexical and discourse constraints, we report results for applying only the lexical constraints (CRF-INF$_{lex}$) as well as results for applying only the discourse constraints (CRF-INF$_{disc}$). (3) PR$_{lex}$: a variant of our PR model which only applies the lexical constraints. For the three-way classification task on the MD dataset, we also implemented the following baselines: (4) VOTEFLIP: a rule-based algorithm that leverages the positive, negative and neutral cues along with the effect of negation to determine the sentence sentiment (Choi and Cardie, 2009). (5) DOCORACLE: assigns each sentence the label of its corresponding document.

[7] Significance test was not conducted over the previous methods as we do not have their results for each fold.

[8] We conducted 10-fold cross-validation on each training fold with the parameter space: $\beta : [0.01, 0.05, 0.1, 0.5, 1.0]$ and $\gamma : [0.1, 0.5, 1.0, 5.0, 10.0]$.

Methods	CR	MD
CRF	81.1	67.0
CRF-inf$_{lex}$	80.9	66.4
CRF-inf$_{disc}$	81.1	67.2
PR$_{lex}$	81.8	69.7
PR	**82.7**	**70.6**
Previous work		
TreeCRF (Nakagawa et al., 2010)	81.4	-
Dropout LR (Wang and Manning, 2013)	82.1	-

Table 2: Accuracy results (%) for supervised sentiment classification (two-way)

	Books	Electronics	Music	Avg
VoteFlip	44.6	45.0	47.8	45.8
DocOracle	53.6	50.5	63.0	55.7
CRF	57.4	57.5	61.8	58.9
CRF-inf$_{lex}$	56.7	56.4	60.4	57.8
CRF-inf$_{disc}$	57.2	57.6	62.1	59.0
PR$_{lex}$	60.3	59.9	63.2	61.1
PR	**61.6**	**61.0**	**64.4**	**62.3**
Previous work				
HCRF	55.9	61.0	58.7	58.5
MEM	59.7	59.6	63.8	61.0

Table 3: Accuracy results (%) for semi-supervised sentiment classification (three-way) on the MD dataset

4.1 Results

We first report results on a binary (positive or negative) sentence-level sentiment classification task. For this task, we used the supervised setting and performed transductive learning for our model. Table 2 shows the accuracy results. We can see

[9] We set λ to 1000 for the lexical constraints and -1000 to the discourse connective constraints in the experiments

	Books pos/neg/neu	**Electronics** pos/neg/neu	**Music** pos/neg/neu
VoteFlip	43/42/47	45/46/44	50/46/46
DocOracle	54/60/49	57/54/42	72/65/52
CRF	47/51/64	60/61/52	67/60/58
CRF-inf$_{lex}$	46/52/63	59/61/50	65/59/57
CRF-inf$_{disc}$	47/51/64	60/61/52	67/61/59
PR$_{lex}$	50/56/66	64/63/53	67/64/59
PR	52/56/68	64/66/53	69/65/60

Table 4: F1 scores for each sentiment category (positive, negative and neutral) for semi-supervised sentiment classification on the MD dataset

that PR significantly outperforms all other baselines in both the CR dataset and the MD dataset (average accuracy across domains is reported). The poor performance of CRF-INF$_{lex}$ indicates that directly applying lexical constraints as hard constraints during inference could only hurt the performance. CRF-INF$_{disc}$ slightly outperforms CRF but the improvement is not significant. In contrast, both PR$_{lex}$ and PR significantly outperform CRF, which implies that incorporating lexical and discourse constraints as posterior constraints is much more effective. The superior performance of PR over PR$_{lex}$ further suggests that the proper use of discourse information can significantly improve accuracy for sentence-level sentiment classification.

We also analyzed the model's performance on a three-way sentiment classification task. By introducing the "neutral" category, the sentiment classification problem becomes harder. Table 4 shows the results in terms of accuracy for each domain in the MD dataset. We can see that both PR and PR$_{lex}$ significantly outperform all other baselines in all domains. The rule-based baseline VOTEFLIP gave the weakest performance because it has no prediction power on sentences with no opinion words. DOCORACLE performs much better than VOTEFLIP and performs especially well on the *Music* domain. This indicates that the document-level sentiment is a very strong indicator of the sentence-level sentiment label. For the CRF baseline and its invariants, we observe a similar performance trend as in the two-way classification task: there is nearly no performance improvement from applying the lexical and discourse-connective-based constraints during CRF inference. In contrast, both PR$_{lex}$ and PR provide substantial improvements over CRF. This confirms that encoding lexical and discourse knowledge as posterior constraints allows the feature-based model to gain additional learning power for sentence-level sentiment prediction. In particular, incorporating discourse constraints leads to consistent improvements to our model. This demonstrates that our modeling of discourse information is effective and that taking into account the discourse context is important for improving sentence-level sentiment analysis. We also compare our results to the previously published results on the same dataset. HCRF (Täckström and McDonald, 2011a) and MEM (Qu et al., 2012) are two state-of-the-art semi-supervised methods for sentence-level sentiment classification. We can see that our best model PR gives the best results in most categories.

Table 4 shows the results in terms of F1 scores for each sentiment category (positive, negative and neutral). We can see that the PR models are able to provide improvements over all the sentiment categories compared to all the baselines in general. We observe that the DOCORACLE baseline provides very strong F1 scores on the positive and negative categories especially in the Books and Music domains, but very poor F1 on the neutral category. This is because it over-predicts the polar sentences in the polar documents, and predicts no polar sentences in the neutral documents. In contrast, our PR models provide more balanced F1 scores among all the sentiment categories. Compared to the CRF baseline and its variants, we found that the PR models can greatly improve the precision of predicting positive and negative sentences, resulting in a significant improvement on the positive/negative F1 scores. However, the improvement on the neutral category is modest. A plausible explanation is that most of our constraints focus on discriminating polar sentences. They can help reduce the errors of misclassifying polar sentences, but the model needs more constraints in order to distinguish neutral sentences from polar sentences. We plan to address this issue in future work.

4.2 Discussion

We analyze the errors to better understand the merits and limitations of the PR model. We found that the PR model is able to correct many CRF errors caused by the lack of labeled data. The first row in Table 5 shows an example of such errors.

Example Sentences	CRF	PR
Example 1: ⟨neg⟩ If I could, I would like to return it or exchange for something better.⟨/neg⟩	⟨neu⟩ ×	✓
Example 2: ⟨neg⟩ Things I wasn't a fan of – the ending was to cutesy for my taste.⟨/neg⟩ ⟨neg⟩ Also, all of the side characters (particularly the mom, vee, and the teacher) were incredibly flat and stereotypical to me.⟨/neg⟩	⟨neu⟩ ⟨pos⟩ ×	✓
Example 3: ⟨neg⟩ I also have excessive noise when I talk and have phone in my pocket while walking.⟨/neg⟩ ⟨neu⟩ But other models are no better.⟨/neu⟩	⟨neg⟩ ⟨pos⟩ ×	⟨neg⟩ ⟨pos⟩ ×

Table 5: Example sentences where PR succeeds and fails to correct the mistakes of CRF

The lexical features *return* and *exchange* may be good indicators of negative sentiment for the sentence. However, with limited labeled data, the CRF learner can only associate very weak sentiment signals to these features. In contrast, the PR model is able to associate stronger sentiment signals to these features by leveraging unlabeled data for indirect supervision. A simple lexicon-based constraint during inference time may also correct this case. However, hard-constraint baselines can hardly improve the performance in general because the contributions of different constraints are not learned and their combination may not lead to better predictions. This is also demonstrated by the limited performance of CRF-INF in our experiments.

We also found that the discourse constraints play an important role in improving the sentiment prediction. The lexical constraints alone are often not sufficient since their coverage is limited by the sentiment lexicon and they can only constrain sentiment locally. On the contrary, discourse constraints are not dependent on sentiment lexicons, and more importantly, they can provide sentiment preferences on multiple sentences at the same time. When combining discourse constraints with features from different sentences, the PR model becomes more powerful in disambiguating sentiment. The second example in Table 5 shows that the PR model learned with discourse constraints correctly predicts the sentiment of two sentences where no lexical constraints apply.

However, discourse constraints are not always helpful. One reason is that they do not constrain the neutral sentiment. As a result they could not help disambiguate neutral sentiment from polar sentiment, such as the third example in Table 5. This is also a problem for most of our lexical constraints. In general, it is hard to learn reliable indicators for the neutral sentiment. In the MD dataset, a neutral label may be given because the sentence contains mixed sentiment or no sentiment or it is off-topic. We plan to explore more refined constraints that can deal with the neutral sentiment in future work. Another limitation of the discourse constraints is that they could be affected by the errors of the discourse parser and the coreference resolver. A potential way to address this issue is to learn discourse constraints jointly with sentiment. We plan to study this in future research.

5 Conclusion

In this paper, we propose a context-aware approach for learning sentence-level sentiment. Our approach incorporates intuitive lexical and discourse knowledge as expressive constraints while training a conditional random field model via posterior regularization. We explore a rich set of context-aware constraints at both intra- and inter-sentential levels, and demonstrate their effectiveness in the analysis of sentence-level sentiment. While we focus on the sentence-level task, our approach can be easily extended to handle sentiment analysis at finer levels of granularity. Our experiments show that our model achieves better accuracy than existing supervised and semi-supervised models for the sentence-level sentiment classification task.

Acknowledgments

This work was supported in part by DARPA-BAA-12-47 DEFT grant #12475008 and NSF grant BCS-0904822. We thank Igor Labutov for helpful discussion and suggestions; Oscar Täckström and Lizhen Qu for providing their Amazon review datasets; and the anonymous reviewers for helpful comments and suggestions.

References

Kedar Bellare, Gregory Druck, and Andrew McCallum. 2009. Alternating projections for learning

with expectation constraints. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 43–50. AUAI Press.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801. Association for Computational Linguistics.

Yejin Choi and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 590–598. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2010. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the ACL-IJCNLP*, pages 369–377.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 99:2001–2049.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

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.

Dan Jurafsky, James H Martin, Andrew Kehler, Keith Vander Linden, and Nigel Ward. 2000. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, volume 2. MIT Press.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *To Appear in Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules.

Yi Mao and Guy Lebanon. 2007. Isotonic conditional random fields and local sentiment flow. *Advances in neural information processing systems*, 19:961.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 432.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. *Opinion mining and sentiment analysis*. Now Pub.

Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Lizhen Qu, Rainer Gemulla, and Gerhard Weikum. 2012. A weakly supervised model for sentence-level semantic orientation analysis with multiple experts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 149–159. Association for Computational Linguistics.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 801–808. Association for Computational Linguistics.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 170–179. Association for Computational Linguistics.

Oscar Täckström and Ryan McDonald. 2011a. Discovering fine-grained sentiment with latent variable structured prediction models. In *Advances in Information Retrieval*, pages 368–374. Springer.

Oscar Täckström and Ryan McDonald. 2011b. Semi-supervised latent variable models for sentence-level sentiment analysis.

Rakshit Trivedi and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings of NAACL-HLT*, pages 808–813.

Sida Wang and Christopher Manning. 2013. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 118–126.

Qi Zhang, Jin Qian, Huan Chen, Jihua Kang, and Xuanjing Huang. 2013. Discourse level explanatory relation extraction from product reviews using first-order logic.

Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171. Association for Computational Linguistics.