Discovering fine-grained sentiment with latent variable structured prediction models

Oscar Täckström*[1,2] and Ryan McDonald[3]

Swedish Institute of Computer Science[1]
Dept. of Linguistics and Philology, Uppsala University[2]
oscar@sics.se
Google, Inc.[3]
ryanmcd@google.com

Abstract. In this paper we investigate the use of latent variable structured prediction models for fine-grained sentiment analysis in the common situation where only coarse-grained supervision is available. Specifically, we show how sentence-level sentiment labels can be effectively learned from document-level supervision using hidden conditional random fields (HCRFs) [10]. Experiments show that this technique reduces sentence classification errors by 22% relative to using a lexicon and 13% relative to machine-learning baselines.[1]

1 Introduction

Determining the sentiment of a fragment of text is a central task in the field of opinion classification and retrieval [8]. Most research in this area can be categorized into one of two categories: lexicon or machine-learning centric. In the former, large lists of phrases are constructed manually or automatically indicating the polarity of each phrase in the list. This is typically done by exploiting common patterns in language [2], lexical resources such as WordNet or thesauri [4], or via distributional similarity [11]. The latter approach – machine-learning centric – builds statistical text classification models based on labeled data, often obtained via consumer reviews that have been tagged with an associated star-rating, e.g., [9]. Both approaches have their strengths and weaknesses. Systems that rely on lexicons can analyze text at all levels, including fine-grained sentence, clause and phrase levels, which is fundamental to building user-facing technologies such as faceted opinion search and summarization [3]. However, lexicons are typically deployed independent of the context in which mentions occur, often making them brittle, especially in the face of domain shift and complex syntactic constructions [12]. The machine-learning approach, on the other hand, can be trained on the millions of labeled consumer reviews that exist on review aggregation websites, but the supervised learning signal is often at too coarse a level.

* Part of this work was performed while the author was an intern at Google, Inc. and part was funded by the Swedish National Graduate School of Language Technology (GSLT).

[1] A longer version of this paper containing comprehensive descriptions of the models and experiments is available at: http://soda.swedish-ict.se/4058.

We propose a model that learns to analyze fine-grained sentiment strictly from coarse annotations. Such a model can leverage the plethora of labeled documents from multiple domains available on the web. In particular, we focus on sentence level sentiment (or polarity) analysis. As input, the system expects a sentence segmented document and outputs the corresponding sentence labels. The model we present is based on hidden conditional random fields (HCRFs) [10], a well-studied latent variable structured learning model that has been used previously in speech and vision. We show that this model naturally fits the task and can reduce fine-grained classification errors by over 20%.

Latent-variable structured learning models have been investigated recently in the context of sentiment analysis. Nakagawa et al. [7] presented a sentence level model with latent variables placed at the nodes from the syntactic dependency tree of the sentence. Yessenalina et al. [13], presented a document level model with binary latent variables over sentences. In both these models, the primary goal was to improve the performance on the supervised annotated signal. This study inverts the evaluation and attempts to assess the accuracy of the latent structure induced from the observed coarse signal. In fact, one could argue that learning fine-grained sentiment from document level labels is the more relevant question for multiple reasons as 1) document level annotations are the most common naturally observed sentiment signal, e.g., star-rated consumer reviews, and 2) document level sentiment analysis is too coarse for most applications, especially those that rely on aggregation and summarization across fine-grained topics [3].

2 A conditional latent variable model of fine-grained sentiment

The distribution of sentences in documents from our data (Table 2) suggests that documents contain at least one dominant class, even though they do not have uniform sentiment. Specifically, positive (negative) documents primarily consist of positive (negative) sentences as well as a significant number of neutral sentences and a small amount of negative (positive) sentences. This observation suggests that we would like a model where sentence level classifications are 1) correlated with the observed document label, but 2) have the flexibility to disagree when contextual evidence suggests otherwise.

To build such a model, we start with the supervised fine-to-coarse sentiment model described in [6]. Let d be a document consisting of n sentences, $\boldsymbol{s} = (s_i)_{i=1}^n$. We denote by $\boldsymbol{y}^d = (y^d, \boldsymbol{y}^s)$ random variables that include the document level sentiment, y^d, and the sequence of sentence level sentiment, $\boldsymbol{y}^s = (y_i^s)_{i=1}^n$. All random variables take values in $\{\text{POS}, \text{NEG}, \text{NEU}\}$ for positive, negative and neutral sentiment respectively. We hypothesize that there is a sequential relationship between sentence sentiment and that the document sentiment is influenced by all sentences (and vice versa). Figure 1a shows an undirected graphical model reflecting this idea. A first order Markov property is assumed, according to which each sentence variable y_i^s is independent of all other variables, conditioned on the document variable y^d and its adjacent sentences, y_{i-1}^s/y_{i+1}^s. It was shown in [6] that when trained with fully supervised data, this model can increase both sentence and document level prediction accuracies.

We are interested in the common case where document labels are available during training, e.g., from star-rated reviews, but sentence labels are not. A modification to Figure 1a is to treat all sentence labels as unobserved as shown in Figure 1b. When the

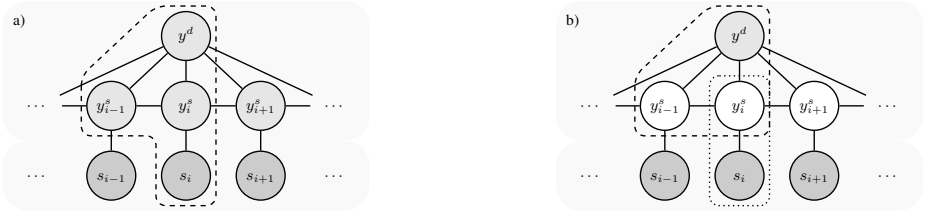

Fig. 1. a) Outline of graphical model from [6]. b) Identical model with latent sentence level states. Dark grey nodes are observed variables and white nodes are unobserved. Light grey nodes are observed at training time. Dashed and dotted regions indicate the maximal cliques at position i.

underlying model from Figure 1a is a conditional random field (CRF) [5], the model in Figure 1b is often referred to as a *hidden* conditional random field (HCRF) [10]. HCRFs are appropriate when there is a strong correlation between the observed coarse label and the unobserved fine-grained variables, which Table 2 indicates is true for our task.

In the CRF model just outlined, the distribution of the random variables $\boldsymbol{y}^d = (y^d, \boldsymbol{y}^s)$, conditioned on input sentences $\boldsymbol{s}$, belongs to the exponential family and is written $p_\theta(y^d, \boldsymbol{y}^s|\boldsymbol{s}) = \exp\left\{\langle\phi(y^d, \boldsymbol{y}^d, \boldsymbol{s}), \theta\rangle - A_\theta(\boldsymbol{s})\right\}$, where θ is a vector of model parameters and $\phi(\cdot)$ is a vector valued feature function, which by the independence assumptions of the graphical models outlined in Figure 1, factorizes as $\phi(y^d, \boldsymbol{y}^s, \boldsymbol{s}) = \oplus_{i=1}^{n}\phi(y^d, y_i^s, y_{i-1}^s, \boldsymbol{s})$, where $\oplus$ indicates vector summation. The log-partition function $A_\theta(\boldsymbol{s})$ is a normalization constant, which ensures that $p_\theta(y^d, \boldsymbol{y}^s|\boldsymbol{s})$ is a proper probability distribution. In an HCRF, the conditional probability of the observed variables is obtained by marginalizing over the posited hidden variables: $p_\theta(y^d|\boldsymbol{s}) = \sum_{\boldsymbol{y}^s} p_\theta(y^d, \boldsymbol{y}^s|\boldsymbol{s})$. As indicated in Figure 1b, there are two maximal cliques at each position i: one involving only the sentence s_i and its corresponding latent variable y_i^s and one involving the consecutive latent variables y_i^s, y_{i-1}^s and the document variable y^d. The assignment of the document variable y^d is thus independent of the input $\boldsymbol{s}$, conditioned on the sequence of latent sentence variables $\boldsymbol{y}^s$. This is in contrast to the original fine-to-coarse model, in which the document variable depends directly on the sentence variables as well as the input [6]. This distinction is important for learning predictive latent variables as it creates a bottleneck between the input sentences and the document label. When we allow the document label to be directly dependent on the input, we observe a substantial drop in sentence level prediction performance.

The feature function at position i is the sum of the feature functions for each clique at that position, that is $\phi(y^d, y_i^s, y_{i-1}^s, \boldsymbol{s}) = \phi(y^d, y_i^s, y_{i-1}^s) \oplus \phi(y_i^s, s_i)$. The features of the clique (y_i^s, s_i) include the identities of the tokens in s_i, the identities and polarities of tokens in s_i that match the polarity lexicon described in [12], and the corresponding number of positive and negative tokens in s_i. Features for (y^d, y_i^s, y_{i-1}^s)-cliques only involve various combinations of the document and sentence sentiment variables.

Typically, when training HCRFs, we find the MAP estimate of the parameters with respect to the marginal conditional log-likelihood of observed variables, assuming a Normal prior, $p(\theta) \sim \mathcal{N}(0, \sigma^2)$. However, early experiments indicated that using *hard* estimation gave slightly better performance. Let $D = \{(\boldsymbol{s}_j, y_j^d)\}_{j=1}^m$ be a training set of document / document-label pairs. In hard HCRF estimation we seek parameters θ such that:

$$\underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{|D|} \log p_\theta(y_j^d, \hat{\boldsymbol{y}}_j^s | \boldsymbol{s}_j) - \frac{\left\|\theta^2\right\|}{2\sigma^2}\,, \quad (1) \qquad \text{with: } \hat{\boldsymbol{y}}_j^s = \underset{\boldsymbol{y}^s}{\operatorname{argmax}}\, p_\theta(y_j^d, \boldsymbol{y}^s | \boldsymbol{s}_j)\,. \quad (2)$$

We find the parameters θ that maximizes (1) by stochastic gradient ascent for 75 iterations, with a fixed step size, η. Contrary to a fully observed CRF, equation (1) is not concave. Still, for the studied data set, results were stable simply by initializing θ to the zero vector. Equation (2) is also used in predicting the optimal assignment of $(y^d, \boldsymbol{y}^s)$. An efficient solution to (2) is provided by the Viterbi algorithm as described in [6].

3 Experiments

For our experiments we constructed a large balanced corpus of consumer reviews from a range of domains. A training set was created by sampling a total of 143,580 positive, negative and neutral reviews from five different domains: *books*, *dvds*, *electronics*, *music* and *videogames*. Document sentiment labels were obtained by labeling one and two star reviews as negative (NEG), three star reviews as neutral (NEU), and four and five star reviews as positive (POS). The total number of sentences is roughly 1.5 million. To study the impact of the training set size, additional training sets, denoted *S* and *M*, were created by sampling 1,500 and 15,000 documents from the full training set, denoted *L*. A smaller separate test set of 294 reviews was constructed by the same procedure. This set consists of 97 positive, 98 neutral and 99 negative reviews. Two annotators marked each test set review sentence as having positive (POS), negative (NEG) or neutral (NEU) sentiment. NEU was used for sentences that are either truly neutral in opinion, have mixed positive/negative sentiment or contain no sentiment. Annotation statistics can be found in Table 1, while Table 2 shows the distribution of sentence level sentiment for each document sentiment category. Overall raw inter-annotator agreement was 86% with a Cohen's κ value of 0.79.[2]

We compare the HCRF model to two baseline models: 1) VoteFlip, which determines the polarity of a sentence with the vote-flip algorithm [1], using the polarity lexicon from [12], and 2) Document as Sentence (DaS), which trains a document classifier on the coarse-labeled training data, but applies it to sentences independently at test time. In order to make the underlying statistical models the same, DaS was parameterized as a log-linear model with a Normal prior distribution and stochastic gradient ascent was used to find the maximum a posteriori point estimate of the parameters. We also measured the benefit of observing the document label at test time. This is a common scenario in, e.g., consumer-review aggregation [3]. The baseline of predicting all sentences with the observed document label is denoted DocOracle. Ten runs were performed with different random seeds and results where gathered by hierarchically bootstrapping medians and confidence intervals, which where also used to test statistical significance.

Table 3 shows results for each model in terms of sentence and document accuracy as well as F_1-scores for each sentence sentiment category. When using enough training data, the HCRF significantly outperforms all baselines in terms of sentence-level accuracy with

[2] The test set is available at `http://www.sics.se/people/oscar/datasets`.

	POS	NEG	NEU	Total	POS	NEG	NEU	Total
Books	19	20	20	59	160	195	384	739
Dvds	19	20	20	59	164	264	371	799
Electronics	19	19	19	57	161	240	227	628
Music	20	20	19	59	183	179	276	638
Videogames	20	20	20	60	255	442	335	1,032
Total	97	99	98	294	923	1,320	1,593	3,836

Table 1. Number of documents per category (left) and number of sentences per category (right) in the test set for each domain.

	POS	NEG	NEU
POS	0.53	0.08	0.39
NEG	0.05	0.62	0.33
NEU	0.14	0.35	0.51

Table 2. Distribution of sentence labels (columns) in documents by their labels (rows) in the test data.

	Sentence acc. (S)	Sentence acc. (M)	Sentence acc. (L)	Sentence F_1 (L)	Document acc. (L)
VoteFlip	41.5 (-1.8, 1.8)	41.5 (-1.8, 1.8)	41.5 (-1.8, 1.8)	45.7 / 48.9 / 28.0	–
DaS	43.8 (-0.9, 0.8)	46.8 (-0.6, 0.7)	47.5 (-0.8, 0.7)	52.1 / 54.3 / 36.0	66.6 (-2.4, 2.2)
HCRF	43.0 (-1.2, 1.3)	**49.1 (-1.4, 1.5)**	**54.4 (-1.0, 1.0)**	**57.8 / 58.8 / 48.5**	64.6 (-2.0, 2.1)
DocOracle	54.8 (-3.0, 3.1)	54.8 (-3.0, 3.1)	54.8 (-3.0, 3.1)	61.1 / 58.5 / 47.0	–
HCRF (obs.)	48.6 (-1.6, 1.4)	54.3 (-1.9, 1.8)	**58.4 (-0.8, 0.7)**	62.0 / **62.3 / 53.2**	–

Table 3. Median and 95% conf. interval for *S*, *M* and *L* training sets. Median F_1-scores for POS/NEG/NEU sentence categories. Bold: significantly better than comparable baselines, $p < 0.05$.

quite a wide margin. Errors are reduced by 22% relative to the lexicon approach and by 13% compared to the machine learning baseline. When document labels are provided at test time, the corresponding reductions are 29% and 21%. In the latter case the reduction compared to the strong DocOracle baseline is only 8%, however, the probabilistic predictions of the HCRF are more useful than the fixed baseline as illustrated by Figure 2. In terms of document accuracy, DaS seems to slightly outperform the HCRF. This is contrary to the results reported in [13], where sentence-level latent variables improved document predictions. However, our model is restricted when it comes to document classification, since document variables are conditionally independent of the input. We observe from Table 3 that all models perform best on positive and negative sentences, with a sharp drop for neutral sentences, though the drop is substantially less for the HCRF. This is not surprising, as neutral documents are bad proxies for sentence sentiment, as can be seen from the sentence sentiment distributions per document category in Table 2.

Adding more training data improves all models and starting with the medium data set, the HCRF outperforms DaS. While the improvement from additional training data is relatively small for DaS, the improvement is substantial for the HCRF. We expect the benefit of using latent variables to increase further with more training data. Finally, Figure 2 shows sentence-level precision–recall curves for the HCRF, with and without observed document label, and DaS, together with the fixed points of VoteFlip and DocOracle. Label probabilities were calculated using the marginal probability of each label at each sentence position. The HCRF dominates at nearly all levels of precision/recall, especially for positive sentences.

Although the results for all systems may seem low, they are comparable with those in [6], which was a fully supervised model and evaluated with neutral documents excluded. In fact, the primary reason for the low scores presented here is the inclusion of neutral documents and sentences. Additional experiments with negative documents excluded showed that the HCRF achieves a document accuracy of 88.4%, which is on par with reported state-of-the-art document accuracies for the two-class task [7, 13]. Furthermore, the annotator agreement of 86% can be viewed as an upper bound on sentence accuracy.

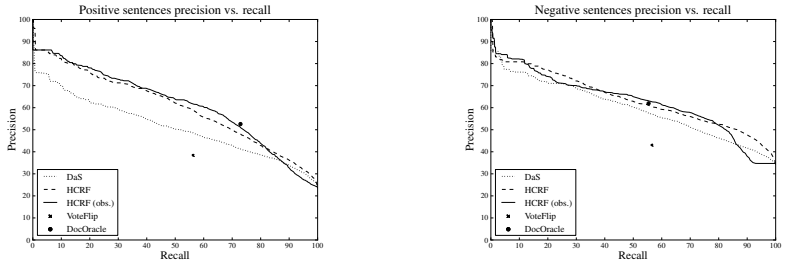

Fig. 2. Interpolated precision-recall curves for positive and negative sentence level sentiment.

4 Conclusions

In this paper we showed that latent variable structured prediction models can effectively learn fine-grained sentiment from coarse-grained supervision. Empirically, reductions in error of up to 20% were observed relative to both lexicon-based and machine-learning baselines. In the common case when document labels are available at test time as well, we observed error reductions close to 30% and over 20%, respectively, relative to the same baselines. In the latter case, our model reduces errors relative to the strongest baseline with 8%. The model we employed, a hidden conditional random field, leaves open a number of further avenues for investigating weak prior knowledge in fine-grained sentiment analysis, most notably semi-supervised learning when small samples of data annotated with fine-grained information are available.

References

1. Y. Choi and C. Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proc. EMNLP*, 2009.
2. V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proc. EACL*, 1997.
3. M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. KDD*, 2004.
4. S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proc. COLING*, 2004.
5. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001.
6. R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proc. ACL*, 2007.
7. T. Nakagawa, K. Inui, and S. Kurohashi. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. In *Proc. NAACL*, 2010.
8. B. Pang and L. Lee. *Opinion mining and sentiment analysis*. Now Publishers, 2008.
9. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. EMNLP*, 2002.
10. A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.
11. P. Turney. Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews. In *Proc. ACL*, 2002.
12. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. EMNLP*, 2005.
13. A. Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *Proc. EMNLP*, 2010.