Sentiment analysis based on probabilistic models using inter-sentence information

Kugatsu Sadamitsu †, Satoshi Sekine ‡, Mikio Yamamoto †

†Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tenoudai, Tsukuba science city, Ibaraki 305-8573, JAPAN
{sadamitsu@mibel.cs,myama@cs}.tsukuba.ac.jp
‡Computer Science Department, New York University
715 Broadway, 7th floor New York, NY 10003, USA
sekine@cs.nyu.edu

Abstract

This paper proposes a new method of the sentiment analysis utilizing inter-sentence structures especially for coping with reversal phenomenon of word polarity such as quotation of other's opinions on an opposite side. We model these phenomenon using Hidden Conditional Random Fields(HCRFs) with three kinds of features: transition features, polarity features and reversal (of polarity) features. Polarity features and reversal features are doubly added to each word, and each weight of the features are trained by the common structure of positive and negative corpus in, for example, assuming that reversal phenomenon occured for the same reason (features) in both polarity corpus. Our method achieved better accuracy than the Naive Bayes method and as good as SVMs.

1. Introduction

Recently, many people express opinions on the WWW, using blogs and online customer reviews. It results in enormous amount of texts including subjective opinions or sentiments. Such information is a valuable for consumers and companies, and there are demands for the efficient analysis of the information. One of the popular analysis of the information is to find if the opinion is positive or negative regarding the target item or service. It is called "sentiment classification"(Turney, 2002)(Pang and Lee, 2002) which is a subject of growing interest. Many past sentiment classification studies only use word-level information in accordance with the polarity of the word. This paper aims to improve the accuracy of sentiment classification by using Hidden Conditional Random Fields (HCRFs) to construct a sentence-by-sentence basis model, thereby capturing the global information that was not captured using the conventional word-by-word model. To be more precise, the chain structure of sentences is modeled using HCRFs that regard sentences as direct output symbols. We will describe the new models achieved better accuracy than the Naive Bayes method and as good as SVMs.

2. Sentiment classification using inter-sentence information

2.1. Reversal of polarity in the sentence level structures

It is often the case that information in inter-sentence should be included in order to make the accurate analysis. Consider the following review from Amazon.com[1].

An example of review

> I like this razor very much. It gives a close shave without damaging the skin. However, the only drawback is that you must constantly be holding in the "On" button. My last razor had a button that when it was switched on, it stayed on.

This is a positive review about the razor, although there are also negative sentences which partially complain about it. In this paper, we hypothesize that this partial reversal of polarity occurs on a sentence-by-sentence basis as the first-order approximate, and such the sentence-level structures are modeled by sequential models such as the HCRFs that regard a sentence as the output unit.

2.2. Modeling of inter-sentence structure using HMMs

When hypothesizing that each sentence has a certain hidden class (e.g. "quotation" class or "evaluation for a different target" class) and the transition of the class makes up structure of the document, it is natural to use HMMs which regard the sentence itself as an output symbol. Moore type sentence based HMMs are considered here for a sentence $\boldsymbol{s}$ that is represented as a set of words. Probability P_{HMM} of a document d using sentence based HMMs is defined as follows.

$$\begin{aligned} P_{HMM}(d|\boldsymbol{a},\boldsymbol{b}) &= \sum_{\boldsymbol{q}_1^T}\prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{s}_t) \\ &= \sum_{\boldsymbol{q}_1^T}\prod_{t=1}^{T} a_{q_{t-1}q_t} \prod_{n=1}^{|\boldsymbol{s}_t|} b_{q_t}(w_{tn}) \quad (1) \end{aligned}$$

Where $\boldsymbol{s}_t$ means t-th sentence in the document d, T means the number of sentences in the document d, w_{tn} means n-th word in the sentence $\boldsymbol{s}_t$, and q_t means the HMMs state of the sentence $\boldsymbol{s}_t$. Furthermore, $\boldsymbol{a}, \boldsymbol{b}$ are model parameters, thereby $a_{q_{t-1}q_t}$ represents the transition probability from

[1] http://www.amazon.com

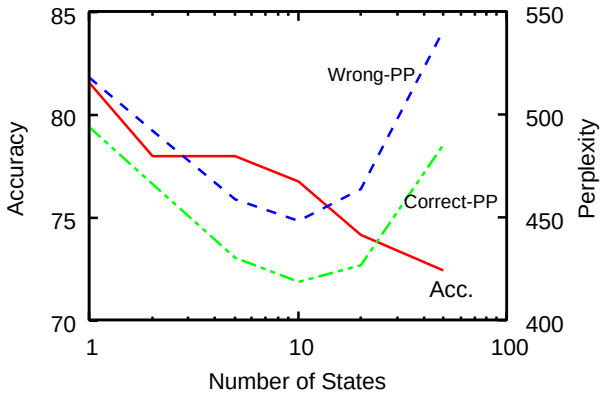

Figure 1: the accuracy of sentiment classification using HMMs

state q_{t-1} to q_t, and $b_{q_t}(w_{tn})$ means the output probability of word w_{tn} in state q_t.

Although HMMs are quite natural and powerful to model hidden class structures in a document, the Maximum Likelihood Estimation (MLE) training of HMM does not always improve classification accuracy appropriately. Because, each polarity model is separately estimated by MLE training using the positive and negative document corpus respectively and the objective function based on likelihood doesn't directly connect to accuracy. In fact, Figure 1 shows that the more the number of states of HMM (the more complex), the lower the accuracy of sentiment classification in our preliminary experiment in which we used the above HMM model and MLE training method based on the EM algorithm. Also Figure 1 shows the relationship between perplexity and classification accuracy for HMMs when varying the number of states, wherein HMMs were trained by MLE in regard to each of positive and negative reviews at Amazon.com. In this figure Correct-PP indicates perplexity for the correct polarity label (positive for the positive model, negative for the negative model) while Wrong-PP indicates perplexity for the wrong polarity label. The greater the difference in perplexity between the correct and wrong labels the more likely classification is advantageously affected. However, as the number of states increases, a decrease is observed in the perplexity of both the correct and wrong labels until the 10-states, and the difference is mostly constant with classification accuracy decreasing. This is caused by the fact that HMMs are generative models and are not discriminative models.

3. Modeling of inter-sentence structure using HCRFs

3.1. Overview of HCRFs

If we have the corpus with sentence tags, we can use Conditional Random Fields (CRFs) (Lafferty et al., 2001) as an extension of the discriminative model of HMMs. CRFs are log-linear models represented as follows, and the features can be designed more arbitrarily than those of HMMs.

$$P_{CRF}(\boldsymbol{q}|d;\boldsymbol{\lambda}) = \frac{\exp\{\boldsymbol{\lambda}\cdot f(\boldsymbol{q},d)\}}{\sum_{\boldsymbol{q}}\exp\{\boldsymbol{\lambda}\cdot f(\boldsymbol{q},d)\}} \quad (2)$$

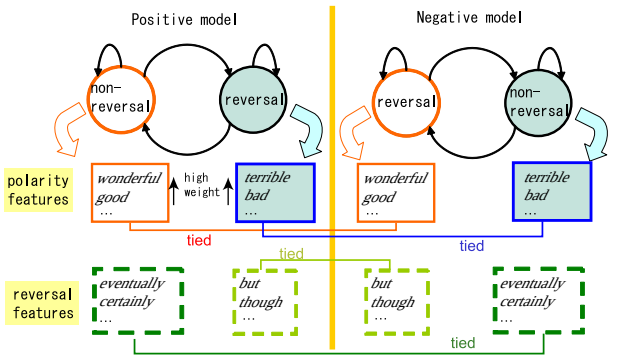

Figure 2: The structures of HCRFs for sentiment classification (assumed each features weight are ideal value)

λ is the parameter vector and $f(\boldsymbol{q},d)$ is a feature vector. However, our purpose is not estimating the series label $\boldsymbol{q}$ of each sentence but rather estimating the document polarity label. HCRFs(Quattoni et al., 2004)(Gunawardana et al., 2005) are appropriate discriminative model for our purpose. HCRFs are obtained by the relative ratio of each polarity label of the denominator which is the sum of the entire state series, and it is evantually defined as follows.

$$P_{HCRF}(\phi|d;\boldsymbol{\lambda}) = \frac{\sum_{\boldsymbol{q}}\exp\{\boldsymbol{\lambda}\cdot f(\phi,\boldsymbol{q},d)\}}{\sum_{\phi'}\sum_{\boldsymbol{q}}\exp\{\boldsymbol{\lambda}\cdot f(\phi',\boldsymbol{q},d)\}} \quad (3)$$

ϕ is the binary label of document polarity either positive or negative. The biggest difference from HMMs are that the conditional probability $P_{HCRF}(\phi|d;\lambda)$ is directly estimated, therefore HCRFs can model inter-sentence structure more appropriately than HMMs in the perspective of classification. Although the number of hidden states, which means the number of classes of the sentences, can be arbitrarily set in the same manner as HMMs, HCRFs under the two states are hereinafter considered because this paper particularly focuses on whether or not a reversal of polarity occurs.

3.2. HCRFs for the sentiment classification

Regarding the advantages of HCRFs, other than the fact that they are discriminative models, another point is that designing the features can be comparatively arbitraly done. In this paper, since only two states of reversal and non-reversal are modeled, words are to be doubly provided with a role as a feature that affects reversal in addition to the role as a polarity feature. However, we cannot simply add words as new features for polarity reversion into the model, because all words were already used for polarity features and it results in merely increasing the number of features that have the same meaning. So we assumed that the relationship between words and polarity reversion is general over the polarity of the document, then we added new features of words tied up to both positive and negative models. Our model is illustrated in Figure 2 and formulated as follows.

$$\begin{aligned} P_{HCRF}(\phi|d;\lambda) &= \frac{1}{Z}\sum_{\boldsymbol{q}}\prod_{t}\exp\{\lambda^{trans}_{q_{t-1},q_t,\phi} \quad (4)\\ &+ \sum_{w\in\boldsymbol{s}_t}(\lambda^{pn}_{w,\gamma(\phi,q_t)} + \lambda^{rev}_{w,q_t})\} \end{aligned}$$

Z is the normalization term. $\lambda^{trans}, \lambda^{pn}$ and λ^{rev} are weighted parameters respectively denoting transition feature, polarity feature, and reversal feature. $\gamma(\phi, q_t)$ returns ϕ, when current state q_t is non-reversal and returns $\bar{\phi}$ (opposite label of ϕ) when q_t is reversal. Similarly regarding the transition feature, although features tied to positive and negative models, in this paper, the decision was made that training would proceed without restrictions. In the latter experimental section, we will show there are slightly different in the transition parameters between each polarity models.

3.3. Setting of loss function

Training with HCRFs proceeds in the direction of increasing the conditional probability. However, the maximizing conditional probability does not always lead to improving classification accuracy. This section introduces a loss function as a new target function directly related to classification errors of HCRFs. Basically, similar to the method of applying a loss function for the CRF of (Suzuki et al., 2006), loss function F is defined by classification error scale D and the sigmoid function.

$$F(d, \phi^c) = \frac{1}{1 + \exp(-\eta D(d, \phi^c))} \quad (5)$$

$$D(d, \phi^c) = -\log p(\phi^c|d; \boldsymbol{\lambda}) + \log p(\phi^w|d; \boldsymbol{\lambda}) \quad (6)$$

Where ϕ^c is a correct polarity label and ϕ^w is a wrong polarity label, $\boldsymbol{\lambda}$ are model parameters, η is the adjustable parameter with gradient in the loss function, $D(d, \phi^c)$ is the miss-classification-measure of a document d. The assignment of formula (5) will result in cancellation of the regularization terms in formula (3). We apply GPD(General Probablistic Descent)(Juang and Katagiri, 1992) to estimate model parameters $\boldsymbol{\lambda}$ and update the parameters as shown below. The arguments in the functions are omitted in order to be simply.

$$\lambda'^{new} = \lambda'^{old} + \rho \frac{\partial F}{\partial \lambda'^{old}}$$
$$= \lambda'^{old} - \rho \eta F(1 - F) \frac{\partial D}{\partial \lambda'^{old}} \quad (7)$$

3.4. Related works

There are some related works, Barzilay et al. proposed modeling the sentence structures using HMMs (Barzilay and Lee, 2004). Their purpose was capturing the topic drift, on the other hand, our purpose is catching the sentiment drift and distinguishing document polarity. Mao et al. introduced the method using CRFs for predicting local sentiment flow in a document (Mao and Guy, 2007). As we described earlier, CRFs needs the training data tagged to the local sentiment, by contrast, HCRFs can be estimated by unsupervised sequential data that causes mitigation of human cost. The most closely related studies to our models are those made by McDonald et al.(McDonald et al., 2007) and Ikeda et al.(Ikeda et al., 2008). The McDonald et al. study is similar to our models wherein features are sequentially designed to perform the polarity estimation of a document. However it needs to sentences with local sentiment tags in the training data, thus requiring natural extensions to the HCRFs which do not require local sentiment tags for sentences. On the other hand, Ikeda et al. targeted the classification of sentiment sentences instead of sentiment documents, thus sequential models were not used. In addition, it does not use the words which are not included in their sentiment word dictionary. To the contrary, we have the different point of view of regarding all words that have possibility of reversing effect.

4. Experiments

4.1. Experimental Settings

Review data for training and evaluation was obtained from Amazon.com. Reviews at Amazon.com are already attached with ratings, 1 to 5 stars (the greater being the better), made by reviewers, and the experiment used 200 reviews from four ratings(1, 2, 4 and 5 stars) of 17 categories, a total of 13,600 reviews [2]. All category labels were removed, and 10-fold cross validation used wherein ratings of 4 and 5 were grouped as a positive review and ratings of 1 and 2 as a negative review. For this study, focusing on features based on unigram words with document frequency of at least 20, a total of 4,051 words were used as the feature presence (0/1). The review titles and names of the reviewers were not included in the training and evaluation data.
Although GPD is used in the parameter estimation of HCRFs, GPD generally has a strong dependency on the initial values. Therefore the initial values are configured as follows. The initial values of the polarity parameters use the Naive Bayes model as is, and it hypothesizes the following mixed model for the reversal weight parameter.

$$P(d; \boldsymbol{p}) = \prod_t \sum_q \frac{p^{pn}(\boldsymbol{s}_t|\gamma(q, \phi))}{\sum_{q'} p^{pn}(\boldsymbol{s}_t|\gamma(q', \phi))} p^{rev}(\boldsymbol{s}_t|q) \quad (8)$$

The initial value of reversal parameter p^{rev} is obtained with the following formula.

$$p^{rev}(v|q) = \frac{\sum_\phi \sum_d \sum_t \frac{p^{pn}(\boldsymbol{s}_t|\gamma(q,\phi))}{\sum_{q'} p^{pn}(\boldsymbol{s}_t|\gamma(q',\phi))} n_{dtv}}{\sum_{v'} \sum_\phi \sum_d \sum_t \frac{p^{pn}(\boldsymbol{s}_t|\gamma(q,\phi))}{\sum_{q'} p^{pn}(\boldsymbol{s}_t|\gamma(q',\phi))} n_{dtv'}} \quad (9)$$

Transition probabilities are uniformly given as 0.5. When converting all parameters to HCRFs as initial parameter values, logarithms of those shall be used in (Gunawardana et al., 2005). In each MCE training, global averaging was employed so as to avoid over-fitting. Training parameters were fixed as $\eta = 1$ and $\rho = 0.0001$, respectively.

4.2. Sentiment classfiication result

Table 1 shows the experimental results of sentiment classification using HCRFs. The baseline is the Naive Bayes model and SVMs. The SVMs training and evaluation uses the software package of SVM^{light}[3] with a linear kernel. Each column of the HCRFs indicates a result under the following conditions; train: when all parameters are learnt. init: when initial values are directly used as parameters of HCRFs. Our method achieved better accuracy than the

[2] although Amazon offers 24 categories on their top page at present, seven of those categories were excluded from the experiment because of an insufficient number of reviews

[3] http://svmlight.joachims.org/

Table 1: Average 10-fold cross-validation accuracies of sentiment classification using the HCRFs and basic models.

NB	SVMs	HCRFs	
		train	init
81.54	82.56	82.74	70.38

Naive Bayes method and as good as SVMs, it can be said that effective training was achieved. In addition, the example review shown in the section 2.1. is an real review which is improved by using HCRFs. On the other hand, an example which is worsen by using HCRFs is following.

- *This is not a scuba diving watch but fun for snorkling and skin diving. The atomic watch updates well in the Washington, DC area. Mine updated the first night and I didn't even put it in a window as suggested. It has synchronized with NIST Ft. Collins every night since. Everything works as advertised.*

The first sentence in this review includes two reversal words "not" and "but". Unfortunatelly, our model can't capture multiple reversal phenomenon. The more reversal words appeared in the sentence, the higher probability of the reversal (negative) hidden state is estimated. This is an important future tasks for our method.

The table 2 shows the training results of the transition parameter (λ^{trans}). π means first sentence transition, "non-rev." means that the sentence polarity is not reversal of the document polarity and "rev." means the sentence polarity is reversal of the document polarity. Some interesting results can be observed here. First, a higher weight is seen in the case where the transition target is in a non-reversal state in comparison to where it is in a reversal state. This result shows that more sentences stay in the non-reversal state as a human's intuition. Regarding initial transition π the negative model has a higher weight in the reversal state (-0.11). This is a characteristic that only appears in this part of the entire table. The reason for this result is presumed that a negative review would have the tendency of first partially acknowledging a positive opinion or of consulting the opinions of others.

Finally, we show the table 3 which has the characteristic words for the polarity and reversal features. These ranks are calculated with the difference in feature weight values to each other. The table shows the top 5 polarity words and the top 2 and selected from the top 50 reversal words, as the top reversal words are a little noisy, similar to the top 2 words. The numbers in the brackets on the right side of the reversal words denote the rank of the words.

5. Conclusion

This paper proposed a new method of the sentiment classification utilizing inter-sentence structures especially for coping with reversal phenomenon of word polarity. Polarity features and reversal features are doubly added to each word, and each weight of features are trained by common

Table 2: The transition weights of HCRFs (λ^{trans}) .

		transition source		
		π	non-rev.	rev.
transition target (posi.)	non-rev.	0.02	1.27	-0.688
	rev.	-1.3	-1.14	-2.03
transition target (nega.)	non-rev.	-0.28	0.74	-0.929
	rev.	-0.11	-1.4	-4.04

Table 3: The characteristic polarity and reversal words.

polarity words (top 5)		reversal words (top 2 and selected from top 50)	
posi.	nega.	non-rev.	rev.
excellent	unusable	unusable(1)	replay(1)
pleased	returned	receipt(2)	lens(2)
pleasantly	disappointing	specially(20)	initially(11)
compliments	disappointed	secondly(34)	vs.(34)
easy	horrible	additionally(50)	e.g.(45)

structure of positive and negative corpus, thereby allowing effective modeling. As the future work we will take into consideration multiple reversal phenomenon, as described in the previous section. Furthermore we are going to apply other parameter training methods, including exponential gradient algorithms(Globerson et al., 2007), more appropriate initial values, the other loss functions, and an method with an increased number of HCRFs states.

6. References

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proc. of the HLT-NAACL 2004 Conf.*, pages 113–120.

A. Globerson, T.Y. Koo, X. Carreras, and M. Collins. 2007. Exponentiated gradient algorithms for log-linear structured prediction. In *Proc. of the 24th ICML*, pages 305–312.

Asela Gunawardana, Milind Mahajan, Alex Acero, and John C. Platt. 2005. Hidden conditional random fields for phone classification. In *Proc. of the INTERSPEECH Conf.*

Daisuke Ikeda, Hiroya Takamura, Lev-Arie Ratinov, and Manabu Okumura. 2008. Learning to shift the polarity of words for sentiment classification. In *Proc. of the IJCNLP-08 Conf.*

B.-H. Juang and S. Katagiri. 1992. Discriminative learning for minimum error classification. In *IEEE Trans. Signal Processing*, volume 40, pages 3043–3054.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning*, pages 282–289.

Yi Mao and Lebanon Guy. 2007. Isotonic conditional ran-

dom fields and local sentiment flow. In *Neural Information Processing Systems*, volume 18.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proc. of the 45th ACL Conf.*, pages 432–439.

Bo Pang and Lillian Lee. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of the EMNLP Conf.*, pages 76–86.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2004. Conditional random fields for object recognition. In *Neural Information Processing Systems*, volume 17.

Jun Suzuki, Erik McDermott, and Hideki Isozaki. 2006. Training conditional random fields with multivariate evaluation measures. In *Proc. of the 21st COLING and 44th ACL Conf.*, pages 217–224.

P. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th ACL Conf.*, pages 417–424.