Adding Redundant Features for CRFs-based Sentence Sentiment Classification

Jun Zhao, Kang Liu, Gen Wang

National Laboratory of Pattern Recognition

Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

{jzhao, kliu, gwang}@nlpr.ia.ac.cn

Abstract

In this paper, we present a novel method based on CRFs in response to the two special characteristics of “contextual dependency” and “label redundancy” in sentence sentiment classification. We try to capture the contextual constraints on sentence sentiment using CRFs. Through introducing redundant labels into the original sentimental label set and organizing all labels into a hierarchy, our method can add redundant features into training for capturing the label redundancy. The experimental results prove that our method outperforms the traditional methods like NB, SVM, MaxEnt and standard chain CRFs. In comparison with the cascaded model, our method can effectively alleviate the error propagation among different layers and obtain better performance in each layer.

1 Introduction[*]

There are a lot of subjective texts in the web, such as product reviews, movie reviews, news, editorials and blogs, etc. Extracting these subjective texts and analyzing their orientations play significant roles in many applications such as electronic commercial, etc. One of the most important tasks in this field is sentiment classification, which can be performed in several levels: word level, sentence level, passage level, etc. This paper focuses on sentence level sentiment classification.

Commonly, sentiment classification contains three layers of sub-tasks. From upper to lower, (1) Subjective/Objective classification: the subjective texts are extracted from the corpus teeming with both subjective and objective texts. (2) Polarity classification: a subjective text is classified into “positive” or “negative” according to the sentimental expressions in the text. (3) Sentimental strength rating: a subjective text is classified into several grades which reflect the polarity degree of “positive” or “negative”. It is a special multi-class classification problem, where the classes are ordered. In machine learning, this kind of problem is also regarded as an ordinal regression problem (Wei Wu et al. 2005). In this paper, we mainly focus on this problem in sentiment classification.

Sentiment classification in sentence level has its special characteristics compared with traditional text classification tasks. Firstly, the sentiment of each sentence in a discourse is not independent to each other. In other words, the sentiment of each sentence is related to those of other adjacent sentences in the same discourse. The sentiment of a sentence may vary in different contexts. If we detach a sentence from the context, its sentiment may not be inferred correctly. Secondly, there is redundancy among the sentiment classes,

[*] Contact: Jun ZHAO, jzhao@nlpr.ia.ac.cn

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 117–126,
Honolulu, October 2008. ©2008 Association for Computational Linguistics

especially in sentimental strength classes. For example:

"I love the scenario of "No country for old man" very much!!"

"This movie sounds good."

The first sentence is labeled as "highly praised" class and the second one is labeled as "something good" class. Both the sentences express positive sentiment for the movie, but the former expresses stronger emotion than the latter. We can see that both "highly praised" and "something good" belong to an implicit class "positive", which can be regarded as the relation between them. If we add these implicit classes in the label set, the sentiment classes will form a hierarchical structure. For example, "positive" can be regarded as the parent class of "highly praised" and "something good", "subjective" can be regarded as the parent class of "positive" and "negative". This implicit hierarchical structure among labels should not be neglected because it may be beneficial for improving the accuracy of sentiment classification. In the paper, we call this characteristic of sentiment classification as "label redundancy". Unfortunately, in our knowledge most of the current research treats sentiment classification as a traditional multi-classification task or an ordinal regression task, which regard the sentimental classes being independent to each other and each sentence is also independent to the adjacent sentences in the context. In other words, they neglect the contextual information and the redundancy among sentiment classes.

In order to consider the contextual information in the process of the sentence sentiment classification, some research defines contextual features and some uses special graph-based formulation, like (Bo Pang, et al. 2005). In order to consider the label redundancy, one potential solution is to use a cascaded framework which can combine subjective/objective classification, polarity classification and sentimental strength classification together, where the classification results of the preceding step will be the input of the subsequent one. However, the subsequent classification cannot provide constraint and correction to the results of the preceding step, which will lead to the accumulation and propagation of the classification errors. As a result, the performance of sentiment analysis of sentences is often not satisfactory.

This paper focuses on the above two special characteristics of the sentiment classification problem in the sentence level. To the first characteristic, we regard the sentiment classification as a sequence labeling problem and use conditional random field (CRFs) model to capture the relation between two adjacent sentences in the context. To the second characteristic, we propose a novel method based on a CRF model, in which the original task is mapped to a classification on a hierarchical structure, which is formed by the original label set and some additional implicit labels. In the hierarchical classification framework, the relations between the labels can be represented as the additional features in classification. Because these features are related to the original labels but unobserved, we name them as "redundant features" in this paper. They can be used to capture the redundant and hierarchical relation between different sentiment classes. In this way, not only the performance of sentimental strength rating is improved, the accuracies of subjective/objective classification and polarity classification are also improved compared with the traditional sentiment classification method. And in comparison with the cascaded method, the proposed approach can effectively alleviate error propagation. The experimental results on movie reviews prove the validity of our method.

2 Capturing Contextual Influence for Sentiment Classification

For capturing the influence of the contexts to the sentiment of a sentence, we treat original sentiment classification as a sequence labeling problem. We regard the sentiments of all the sentences throughout a paragraph as a sequential flow of sentiments, and we model it using a conditional model. In this paper, we choose Conditional Random Fields (CRFs) (Lafferty et al, 2001) because it has better performance than other sequence labeling tools in most NLP applications.

CRFs are undirected graphical models used to calculate the conditional probability of a set of labels given a set of input variables. We cite the definitions of CRFs in (Lafferty et al, 2001). It defines the conditional probability proportional to the product of potential functions on cliques of the graph,

$$P_{\lambda}(Y \mid X) = \frac{\exp \lambda \cdot F(Y, X)}{Z(X)} \qquad (1)$$

where X is a set of input random variables and Y is a set of random labels. $F(Y, X)$ is an arbitrary feature function over its arguments, λ is a learned weight for each feature function and $Z(X) = \sum_{y} \exp(\lambda \cdot F(Y, X))$.

The training of CRFs is based on Maximum Likelihood Principle (Fei Sha et al. 2003). The log likelihood function is

$$L(\lambda) = \sum_{k} \left[\lambda \cdot F(Y_k, X_k) - \log Z_{\lambda}(X_k) \right]$$

Therefore, Limited-memory BFGS (L-BFGS) algorithm is used to find this nonlinear optimization parameters.

3 Label Redundancy in Sentiment Classification

In this section, we explain the “label redundancy” in sentiment classification mentioned in the first section. We will analyze the effect of the label redundancy on the performance of sentiment classification from the experimental view.

We conduct the experiments of polarity classification and sentimental strength rating on the corpus which will be introduced in section 5 later. The class set is also illustrated in that section. Polarity classification is a three-class classification process, and sentimental strength rating is a five-class classification process. We use first 200 reviews as the training set which contains 6,079 sentences, and other 49 reviews, totally 1,531 sentences, are used as the testing set. Both the three-class classification and the five-class classification use standard CRFs model with the same feature set. The results are shown in Table 1, 2 and 3, where “Answer” denotes the results given by human, “Results” denotes the results given by CRFs model，“Correct” denotes the number of correct samples which is labeled by CRFs model. We use precision, recall and F1 value as the evaluation metrics.

Table 1 gives the result of sentimental strength rating. Table 2 shows the polarity classification results extracted from the results of sentimental strength rating in Table 1. The extraction process is as follows. In the sentimental strength rating results, we combine the sentences with “PP” class and the sentences with “P” class into “Pos” class, and the sentences with “NN” class and the sentences with “N” class into “Neg” class. So the results of five-class classification are transformed into the results of three-class classification. Table 3 is the results of performing polarity classification in the data set by CRFs directly.

Label	Answer	Results	Correct	Precision	Recall	F1
PP	51	67	5	0.0746	0.0980	0.0847
P	166	177	32	0.1808	0.1928	0.1866
Neu	1190	1118	968	0.8658	0.81.34	0.8388
N	105	140	25	0.1786	0.2381	0.2041
NN	19	29	1	0.0345	0.0526	0.0417
Total	1531	1531	1031	0.67.34	0.6734	0.6734

Table 1. Result of Sentimental Strength Rating

Label	Answer	Results	Correct	Precision	Recall	F1
Pos	217	244	79	0.3238	0.3641	0.3427
Neu	1190	1118	968	0.8658	0.8134	0.8388
Neg	124	169	41	0.2426	0.3306	0.2799
Total	1531	1531	1088	0.7106	0.7106	0.7106

Table 2. Result of Polarity Classification Extracted from Table 1.

Label	Answer	Results	Correct	Precision	Recall	F1
Pos	217	300	108	0.3600	0.4977	0.4178
Neu	1190	1101	971	0.8819	0.8160	0.8477
Neg	124	130	40	0.3077	0.3226	0.3150
Total	1531	1531	1119	0.7309	0.7309	0.7309

Table 3. Result of Polarity Classification

From the results we can find the following phenomena.

(1) The corpus is severely unbalanced, the objective sentences take the absolute majority in the corpus, which leads to the poor accuracy for classifying subjective sentences. The experiment in Table 1 puts polarity classification and sentimental strength rating under a unique CRFs model, without considering the redundancy and hierarchical structure between different classes. As a result, the features for polarity classification will usually cover the features for sentimental strength rating. These reasons can explain why there is only one sample labeled as "NN" correctly and only 5 samples labeled as "PP" correctly.

(2) Comparing Table 2 with 3, we can find that, the F1 value of the polarity classification results extracted from sentimental strength rating results is lower than that of directly conducting polarity classification. That is because the redundancy between sentimental strength labels makes the classifier confused to determine the polarity of the sentence. Therefore, we should deal with the sentiment analysis in a hierarchical frame which can consider the redundancy between the different classes and make full use of the subjective and polarity information implicitly contained in sentimental strength classes.

4 Capturing Label Redundancy for CRFs via Adding Redundant Features

As mentioned above, it's important for a classifier to consider the redundancy between different labels. However, from the standard CRFs described in formula (1), we can see that the training of CRFs only maximizes the probabilities of the observed labels Y in the training corpus. Actually, the redundant relation between sentiment labels is unobserved. The standard CRFs still treats each class as an isolated item so that its performance is not satisfied.

In this section, we propose a novel method for sentiment classification, which can capture the redundant relation between sentiment labels through adding redundant features. In the following, we firstly show how to add these redundant features, then illustrate the characteristics of this method. After that, for the sentiment analysis task, the process of feature generation will be presented.

4.1 Adding Redundant Features for CRFs

Adding redundant features has two steps. Firstly, an implicit redundant label set is designed, which can form a multi-layer hierarchical structure together with the original labels. Secondly, in the hierarchical classification framework, the implicit labels, which reflect the relations between the original labels, can be used as redundant features in the training process. We will use the following example to illustrate the first step for sentimental strength rating task.

For the task of sentimental strength rating, the original label set is {"PP (highly praised)", "P (something good)", "Neu (objective description)", "N (something that needs improvement)" and "NN (strong aversion)"}. In order to introduce redundant labels, the 5-class classification task is decomposed into the following three layers shown in Figure 1. The label set in the first layer is {"subjective", "objective"}, The label set in the second layer is for polarity classification {"positive", "objective", "negative"}, and the label set in the third layer is the original set. Actually, the labels in the first and second layers are unobserved redundant labels, which will not be reflected in the final classification result obviously.

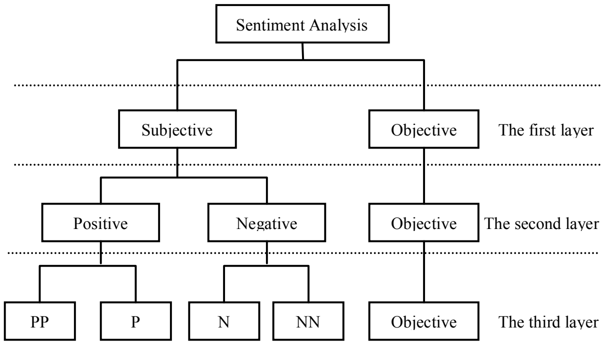

Figure 1. The hierarchical structure of sentimental labels

In the second step, with these redundant labels, some implicit features can be generated for CRFs. So the standard CRFs can be rewritten as follows.

$$P(T \mid X) = \frac{\exp(F(X,T)\cdot\lambda)}{Z_T(X)} = \frac{\exp(\sum_{j=1}^{m} F_j(X,Y_j)\cdot\lambda_j)}{\sum_T \exp(\sum_{j=1}^{m} F_j(X,Y_j)\cdot\lambda_j)} \qquad (2)$$

where $T = (Y_1, Y_2, ...Y_j..., Y_m)$, and Y_j denotes the label sequence in the j^{th} layer. $F_j(X,Y_j)$ denotes the arbitrary feature function in the j^{th} layer.

From the formula (2), we can see that the original label set is rewritten as $T = (Y_1, Y_2, ...Y_j..., Y_m)$, which contains implicit labels in the hierarchical structure shown in Figure 1. The difference between our method and the standard chain CRFs is that we make some implicit redundant features to be active when training. The original feature function $F(Y,X)$ is replaced by $\sum_{j=1}^{m} F_j(X,Y_j)$. We use an example to illustrate the process of feature generation. When a sentence including the word "good" is labeled as "PP", our model not only generate the state feature (good, "PP"), but also two implicit redundant state feature (good, "positive") and (good, "subjective"). Through adding larger-granularity labels "positive" and "negative" into the model, our method can increase the probability of "positive" and decrease the probability of "negative". Furthermore, "P" and "PP" will share the probability gain of "positive", therefore the probability of "P" will be larger than that of "N". For the transition feature, the same strategy is used. Therefore the complexity of its training procedure is $O(M \times N \times \sum_{j}^{m} F_j \times l)$ where M is the number of the training samples, N is the average sentence length, F_j is the average number of activated features in the j^{th} layer, l is the number of the original labels and m is the number of the layers. For the complexity of the decoding procedure, our method has $O(N \times \sum_{j}^{m} F_j \times l)$.

It's worth noting that, (1) transition features are extracted in each layer separately rather than across different layers. For example, feature (good, "subjective", "positive") will never be extracted because "subjective" and "positive" are from different layers; (2) if one sentence is labeled as "Neu", no implicit redundant features will be generated.

4.2 The Characteristics of Our Method

Our method allows that the label sets are dependent and redundant. As a result, it can improve the performance of not only the classifier for the original sentimental strength rating task, but also the classifiers for other tasks in the hierarchical frame, i.e. polarity classification and subjective/objective classification. This kind of dependency and redundancy can lead to two characteristics of the proposed method for sentiment classification compared with traditional methods, such as the cascaded method.

(1) Error-correction: Two dependent tasks in the neighboring layers can correct the errors of each other relying on the inconsistent redundant information. For example, if in the first layer, the features activated by "objective" get larger scores than the features activated by "subjective", and in the second layer the features activated by "positive" get larger scores than the features activated by "objective", then inconsistency emerges. At this time, our method can globally select the label with maximum probability. This characteristic can make up the deficiency of the cascaded method which may induce error propagation.

(2) Differentiating the ordinal relation among sentiment labels: Our method organizes the ordinal sentiment labels into a hierarchy through introducing redundant labels into standard chain CRFs, in this way the degree of classification errors can be controlled. In the different layers of sentiment analysis task, the granularities of classification are different. Therefore, when an observation cannot be correctly labeled on a smaller-granularity label set, our method will use the larger-granularity labels in the upper layer to control the final classification labels.

4.3 Feature Selection in Different Layers

For feature selection, our method selects different features for each layer in the hierarchical frame.

In the top layer of the frame shown in Figure 1, for subjective/objective classification task, we use not only adjectives and the verbs which contain subjective information (e.g., “believe”, “think”) as the features, but also the topic words. The topic words are defined as the nouns or noun phases which frequently appear in the corpus. We believe that some topic words contain subjective information.

In the middle and bottom layers, we not only use the features in the first layer, but also some special features as follows.

(1) The prior orientation scores of the sentiment words: Firstly, a sentiment lexicon is generated by extending the synonymies and antonyms in WordNet[2] from a positive and negative seed list. Then, the positive score and the negative score of a sentiment word are individually accumulated and weighted according to the polarity of its synonymies and antonyms. At last we scale the normalized distance of the two scores into 5 levels, which will be the prior orientation of the word. When there is a negative word, like {not, no, can’t, merely, never, …}, occurring nearby the feature word in the range of 3 words size window, the orientation of this word will be reversed and “NO” will be added in front of the original feature word for creating a new feature word.

(2) Sentence transition features: We consider two types of sentence transition features. The first type is the conjunctions and the adverbs occurring in the beginning of this sentence. These conjunctions and adverbs are included in a word list which is manually selected, like {and, or, but, though, however, generally, contrarily, …}. The second type of the sentence transition feature is the position of the sentence in one review. The reason lies in that: the reviewers often follow some writing patterns, for example some reviewers prefer to concede an opposite factor before expressing his/her real sentiment. Therefore, we divide a review into five parts, and assign each sentence with the serial number of the part which the sentence belongs to.

5 Experiments

5.1 Data and Baselines

In order to evaluate the performance of our method, we conducted experiments on a sentence level annotation corpus obtained from Purdue University, which is also used in (Mao and Lebanon 07). This corpus contains 249 movie reviews and 7,610 sentences totally, which is randomly selected from the Cornell sentence polarity dataset v1.0. Each sentence was hand-labeled with one of five classes: PP (highly praised), P (something good), Neu (objective description), N (something that needs improvement) and NN (strong aversion), which contained the orientation polarity of each sentence. Based on the 5-class manually labeled results mentioned above, we also assigned each sentence with one of three classes: Pos (positive polarity), Neu (objective description), Neg (negative polarity). Data statistics for the corpus are given in Table 4.

Label	Pos		Neu	Neg		Total
	PP	P	Neu	N	NN	
5 classes	383	860	5508	694	165	7610
3 classes	1243		5508	859		7610

Table 4. Data Statistics for Movies Reviews Corpus

There is a problem in the dataset that more than 70% of the sentences are labeled as “Neu” and labels are seriously unbalanced. As a result, the “Neu” label is over-emphasized. For this problem, Mao and Lebanon (2007) made a balanced data set (equal number sentences for different labels) which is sampled in the original corpus. Since randomly sampling sentences from the original corpus will break the intrinsic relationship between two adjacent sentences in the context, we don’t create balanced label data set.

For the evaluation of our method, we choose accuracy as the evaluation metrics and some classical methods as the baselines. They are Naïve Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (MaxEnt) (Kamal Nigam et al. 1999) and standard chain CRFs (Fei et al. 2003). We also regard cascaded-CRFs as our baseline for comparing our method with the cascaded-based method. For NB, we use Laplace smoothing method. For SVM, we use the LibSVM[3] with a linear kernel function[4]. For MaxEnt, we use the implementation in the toolkit *Mallet*[5]. For CRFs,

[2] http://wordnet.princeton.edu/

[3] http://www.csie.ntu.tw/~cjlin/libsvm

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

[5] http://mallet.cs.umass.edu/index.php/Main_Page

Label	NB	SVM	MaxEnt	Standard CRF	Cascaded CRF	Our Method
PP	0.1745	0.2219	0.2055	0.2027	**0.2575**	0.2167
P	0.2049	0.2877	0.2353	0.2536	0.2881	**0.3784**
Neu	0.8083	**0.8685**	0.8161	0.8273	0.8554	0.8269
N	0.2636	0.3014	0.2558	0.2981	0.3092	**0.4204**
NN	0.0976	0.1162	0.1148	0.1379	0.1510	**0.2967**
Total	0.6442	0.6786	0.6652	0.6856	0.7153	**0.7521**

Table 5. The accuracy of Sentimental Strength Rating

Label	NB	SVM	MaxEnt	Standard CRF	Cascaded-CRF	Our Method
Pos	0.4218	0.4743	0.4599	0.4405	0.5122	**0.6008**
Neu	0.8147	0.8375	0.8424	0.8260	**0.8545**	0.8269
Neg	0.3217	0.3632	0.2739	0.3991	0.4067	**0.5481**
Total	0.7054	0.7322	0.7318	0.7327	0.7694	**0.7855**

Table 6．The Results of Polarity Classification

Label	NB	SVM	MaxEnt	Standard CRF	Our Method
Subjective	0.4743	0.5847	0.4872	0.5594	**0.6764**
Objective	0.8170	0.8248	0.8212	**0.8312**	0.8269
Total	0.7238	0.7536	0.7518	0.7561	**0.8018**

Table 7. The accuracy of Subjective/Objective Classification

we use the implementation in Flex-CRFs[6]. We set the iteration number to 120 in the training process of the method based on CRFs. In the cascaded model we set 3 layers for sentimental strength rating, where the first layer is subjective/objective classification, the second layer is polarity classification and the last layer is sentimental strength classification. The upper layer passes the results as the input to the next layer.

5.2 Sentimental Strength Rating

In the first experiment, we evaluate the performance of our method for sentimental strength rating. Experimental results for each method are given in Table 5. We not only give the overall accuracy of each method, but also the performance for each sentimental strength label. All baselines use the same feature space mentioned in section 4.3, which combine all the features in the three layers together, except cascaded CRFs and our method. In cascaded-CRFs and our method, we use different features in different layers mentioned in section 4.3. These results were gathered using 5-fold cross validation with one fold for testing and the other 4 folds for training.

From the results, we can obtain the following conclusions. (1) The three versions of CRFs perform consistently better than Naïve Bayes, SVM and MaxEnt methods. We think that is because CRFs model considers the contextual influence of each sentence. (2) Comparing the performance of cascaded CRFs with that of standard sequence CRFs, we can see that not only the overall accuracy but also the accuracy for each sentimental strength label are improved, where the overall accuracy is increased by 3%. It proves that taking the hierarchical relationship between labels into account is very essential for sentiment classification. The reason is that: the cascaded model performs sentimental strength rating in three hierarchical layers, while standard chain CRFs model treats each label as an independent individual. So the performance of the cascaded model is superior to the standard chain CRFs. (3) The experimental results also show that our method performs better than the Cascaded CRFs. The classification accuracy is improved from 71.53% to 75.21%. We think that is because our method adds the label redundancy among the sentimental strength labels into consideration through adding redundant features into the feature sets, and the three subtasks in the cascaded model are merged into a unified model. So the output result is a global optimal result. In this way, the problem of error propagation in the cascaded frame can be alleviated.

[6] http://flexcrfs.sourceforge.net

5.3 Sentiment Polarity Classification

In the second experiment, we evaluate the performance of our method for sentiment polarity classification. Our method is based on a hierarchical frame, which can perform different tasks in different layers at the same time. For example, it can determine the polarity of sentences when sentimental strength rating is performed. Here, the polarity classification results of our method are extracted from the results of the sentimental strength rating mentioned above. In the sentimental strength rating results, we combine the sentences with PP label and the sentences with P label into one set, and the sentences with NN label and the sentences with N label into one set. So the results of 5-class classification are transformed into the results of 3-class classification. Other methods like NB, SVM, MaxEnt, standard chain CRFs perform 3-class classification directly, and their label sets in the training corpus is {Pos, Neu, Neg}. The parameter setting is the same as sentimental strength rating. For the cascaded-CRFs method, we firstly perform subjective/objective classification, and then determine the polarity of the sentences based on the subjective sentences. The experimental results are given in Table 6.

From the experimental results, we can obtain the following conclusion for sentiment polarity classification, which is similar to the conclusion for sentimental strength rating mentioned in section 5.2. That is both our model and the cascaded model can get better performance than other traditional methods, such as NB, SVM, MaxEnt, etc. But the performance of the cascaded CRFs (76.94%) is lower than that of our method (78.55%). This indicates that because our method exploits the label redundancy in the different layers, it can increase the accuracies of both polarity classification and sentimental strength rating at the same time compared with other methods.

5.4 Subjective/Objective Classification

In the last experiment, we test our method for subjective/objective classification. The subjective/objective label of the data is extracted from its original label like section 5.3. As the same as the experiment for polarity classification, all baselines perform subjective/objective classification directly. It's no need to perform the cascaded-based method because it's a 2-class task. The results of our method are extracted from the results of the sentimental strength rating too. The results are shown in Table 7. From it, we can obtain the similar conclusion, i.e. our method outperforms other methods and has the 80.18% classification accuracy. Our method, which introduces redundant features into training, can increase the accuracies of all tasks in the different layers at the same time compared with other baselines. It proves that considering label redundancy are effective for promoting the performance of a sentimental classifier.

6 Related Works

Recently, many researchers have devoted into the problem of the sentiment classification. Most of researchers focus on how to extract useful textual features (lexical, syntactic, punctuation, etc.) for determining the semantic orientation of the sentences using machine learning algorithm (Bo et al. 2002; Kim and Hovy, 2004; Bo et al. 2005, Hu et al. 2004; Alina et al 2008; Alistair et al 2006). But fewer researchers deal with this problem using CRFs model.

For identifying the subjective sentences, there are several research, like (Wiebe et al, 2005). For polarity classification on sentence level, (Kim and Hovy, 2004) judged the sentiment by classifying a pseudo document composed of synonyms of indicators in one sentence. (Pang and Lee, 04) proposed a semi-supervised machine learning method based on subjectivity detection and minimum-cut in graph.

Cascaded models for sentiment classification were studied by (Pang and Lee, 2005). Their work mainly used the cascaded frame for determining the orientation of a document and the sentences. In that work, an initial model is used to determine the orientation of each sentence firstly, then the top subjective sentences are input into a document - level model to determine the document's orientation.

The CRFs has previously been used for sentiment classification. Those methods based on CRFs are related to our work. (Mao et al, 2007) used a sequential CRFs regression model to measure the polarity of a sentence in order to determine the sentiment flow of the authors in reviews. However, this method must manually select a word set for constraints, where each selected word achieved the highest correlation with the sentiment. The performance of isotonic CRFs is strongly related to the selected word set. (McDonald et al 2007; Ivan et al 2008) proposed a structured model based on CRFs for jointly classifying the sentiment of text at varying levels of granularity. They put the sentence level and document level sentiment analysis in an integrated model and employ the orientation of the document to influence the decision of sentence's orientation. Both the above two methods didn't consider the redundant and hierarchical relation between sentimental strength labels. So their methods cannot get better results for the problem mentioned in this paper.

Another solution to this problem is to use a joint multi-layer model, such as dynamic CRFs, multi-layer CRFs, etc. Such kind of models can treat the three sub-tasks in sentiment classification as a multi-task problem and can use a multi-layer or hierarchical undirected graphic to model the sentiment of sentences. The main difference between our method and theirs is that we consider the problem from the feature representation view. Our method expands the feature set according to the number of layers in the hierarchical frame. So the complexity of its decoding procedure is lower than theirs, for example the complexity of the multi-layer CRFs is $O(N \times F \times \prod_j l_j)$ when decoding and our method only has $O(N \times \sum_j F_j \times l)$, where N is the average sentence length, F_j is the average number of activated features in the j^{th} layer, l is the number of the original labels.

7 Conclusion and Future Work

In the paper, we propose a novel method for sentiment classification based on CRFs in response to the two special characteristics of "contextual dependency" and "label redundancy" in sentence sentiment classification. We try to capture the contextual constraints on the sentence sentiment using CRFs. For capturing the label redundancy among sentiment classes, we generate a hierarchical framework through introducing redundant labels, under which redundant features can be introduced. The experimental results prove that our method outperforms the traditional methods (like NB, SVM, ME and standard chain CRFs). In comparison with cascaded CRFs, our method can effectively alleviate error propagation among different layers and obtain better performance in each layer.

For our future work, we will explore other hierarchical models for sentimental strength rating because the experiments presented in this paper prove this hierarchical frame is effective for ordinal regression. We would expand the idea in this paper into other models, such as Semi-CRFs and Hierarchical-CRFs.

Acknowledgments

The work is supported by the National Natural Science Foundation of China under Grants no. 60673042, the Natural Science Foundation of Beijing under Grants no. 4073043 and the National High Technology Development 863 Program of China under Grants no. 2006AA01Z144.

References

Alina A. and Sabine B. 2008. *When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging*. In *Proc. of ACL-08*

Alistair Kennedy and Diana Inkpen. 2006. *Sentiment Classification of Movie Reviews Using Contextual Valence Shifter*s. *Computational Intelligence*, 22(2), pages 110-125

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. *Thumbs up? Sentiment classification using machine learning techniques*. In *Proceedings of EMNLP 2002,* pp.79-86.

Bo Pang and Lillian Lee. 2004. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* In *Proceedings of ACL 2004*, pp.271-278.

Bo Pang and Lillian Lee. 2005. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales*. In *Proceedings of ACL 2005*, pp.115-124.

Ivan Titov and Ryan McDonald. 2008. *A Joint Model o f Text and Aspect Ratings of Sentiment Summarization.* In *Proceedings of ACL-08*, pages 308-316

Janyce Webie, Theresa Wilson and Claire Cardie. 2005. *Annotating expressions of opinions and emotions in lauguage. Language Resources and Evaluation 2005*

Fei Sha and Fernando Pereira, 2003 *Shallow Parsing with Conditional Random Fields*, In *Proceedings ofHLT-NAACL 2003, Edmonton, Canada*, pp. 213-220.

Kim, S and Edward H. Hovy. 2004. *Determining the Sentiment of Opinions*. In *Proceedings of COLING-04*.

Kamal Nigam, John Lafferty and Andrew McCallum. 1999. *Using Maximum Entropy for Text Classification*. In *Proceedings of IJCAI Workshop on Machine Learning for Information Filtering*, pages 61-67.

J Lafferty, A McCallum, F Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In *Proceedings of ICML-01*, pages 282.289.

L. Zhuang, F. Jing, and X.Y. Zhu. 2006. *Movie review mining and summarization*. In *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM)*, pages 43-50.

M. Hu and B. Liu. 2004a. *Mining and summarizing customer reviews*. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168-177.

Ryan McDonald, Kerry Hannan and Tyler Neylon et al. *Structured Models for Fine-to-Coarse Sentiment Analysis*. In *Proceedings of ACL 2007*, pp. 432-439.

Wei Wu, Zoubin Ghahraman, 2005. *Gaussian Processes for Oridinal Regression*. *The Journal of Machine learning Research*, 2005

Y. Mao and G. Lebanon, 2007. *Isotonic Conditional Random Fields and Local Sentiment Flow. Advances in Neural Information Processing Systems* 19, 2007