Sentence Level Subjectivity and Sentiment Analysis Experiments in NTCIR-7 MOAT Challenge

Lizhen Qu   Cigdem Toprak   Niklas Jakob   Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Computer Science Department, Technische Universität Darmstadt, Hochschulstraße 10
D-64289 Darmstadt, Germany
{qu,c_toprak,njakob,gurevych}@tk.informatik.tu-darmstadt.de

Abstract

This paper describes our supervised approach to the opinionated and the polarity subtasks in the NTCIR-7 MOAT Challenge. We apply a sequential tagging approach at the token level and use the learned token labels in the sentence level classification tasks. In our formal run submissions, we utilized SVM^{hmm} in both tasks with syntactic and lexicon-based features. Additionally, we present our experiments with structural correspondence learning (SCL) for addressing the domain adaptation problem in sentiment analysis. We report experiments on three corpora: MPQA, NTCIR-6 and NTCIR-7, however our formal run submission is trained on MPQA. We reached an F-measure of 0.48 (lenient) in the opinionated and 0.27 (lenient) in the polarity subtasks.

Keywords: *Subjectivity Analysis, Sentiment Analysis, Sequential Tagging, Domain Adaptation.*

1 Introduction

Subjectivity and sentiment analysis, a.k.a. opinion mining, are computational linguistics tasks focusing on the computational treatment of subjectivity, sentiments and opinions in text. Recently, subjectivity and sentiment analysis applications started to gain importance as they support information search and data analysis with an in-depth analysis of subjective content.

More specifically, subjectivity analysis aims at automatically recognizing subjective content, i.e., classifying the content as objective vs. subjective. Sentiment analysis, on the other hand, involves several additional sub-tasks, such as: *(i)* determining the emotional orientation (polarity) of the subjective content, i.e., determining whether the analysed content conveys a positive, negative or neutral attitude towards its target, *(ii)* determining the strength of the polarity, i.e., determining whether it is mildly or strongly positive or negative, *(iii)* determining the targets of the opinions in text, and *(iv)* determining the holders of the opinions in text.

NTCIR Multilingual Opinion Analysis Task (MOAT) 2008 [18] focused on the subjectivity and sentiment analysis in newspaper genre with various subtasks including subjectivity classification, polarity classification, holder and target extraction. We participated in the following subtasks:

1. *opinionated subtask* which required a binary classification of sentences for subjectivity, i.e., determining whether the sentence contains opinions or not;

2. *polarity subtask* which required a ternary classification of sentences or subsentences for polarity, i.e., for the opinionated sentences it required assigning polarities (positive, negative or neutral) to each opinion unit (subsentence containing a distinct opinion).

For both opinionated and polarity subtask we applied a supervised sequential tagging approach. Thereby, we made use of lexical, syntactic and lexicon-based features. We experimented with two different lexicons including SentiWordNet [9] and a list of subjectivity clues from previous works [22, 24]. Our formal run submissions for both subtasks are based on models generated using the MPQA corpus[1] [23] consisting of 535 newswire documents with expression level subjectivity annotations.

In the opinionated subtask, we trained a model which labels each token as an opinion expression, a holder, or none of them. Similarly, in the polarity subtask, our model labels each token with a certain polarity. In both subtasks, we propagate the token level labeling predictions to sentence level classification results using some heuristic rules. Besides our formal submission configurations, we experimented with structural correspondence learning techniques to overcome the domain adaptation problem in sentiment analysis.

[1] Available at http://www.cs.pitt.edu/mpqa/

This paper is organized as follows: Section 2 introduces related work in the fields of subjectivity and sentiment analysis. Section 3 presents our approach to the opinionated and polarity subtasks. Section 4 discusses resulting experimental results on various newspaper corpora. Section 5 presents domain adaptation experiments followed by conclusions in Section 6.

2 Related Work

Related work in subjectivity and sentiment analysis can be categorised based on the granularity of the unit being analyzed. Term level work focuses on determining term subjectivity and polarity using corpus-based methods [11], statistical word association methods [20], methods exploiting the graph structure derived from the lexical semantic relationships in WordNet [12, 10], and methods based on the classification of the term's glosses using seed terms of known polarity [8]. Document level work focuses on supervised [16] and unsupervised [21] classification of reviews utilizing the information from the term level classification. However, for the opinionated and the polarity subtasks, most relevant work includes the work at the expression/sentence level.

Supervised approaches to the expression / sentence level subjectivity classification usually utilize features based on the existence of the precompiled subjectivity clues, these clues may be a list of terms with known polarities or lexico-syntactic extraction patterns. In [22], Wiebe and Riloff train sentence level subjectivity and objectivity classifiers using a subjective term list and subjective/objective extraction patterns as features. They bootstrap the patterns in an extraction-pattern learning and supervised sentence classification cycle. In [25], Yu and Hatzivassiloglou train sentence classifiers using the counts of semantically oriented unigrams and bigrams as features. They also assign polarities to subjective sentences by averaging the semantic orientation of subjectivity clues in a sentence.

The closest works to our approach in both subtasks include expression level classification presented in [24, 6, 4]. In [24], Wilson et al. attempt to disambiguate the polarities of subjective clue instances in context utilizing a rich set of features encoding syntactic and word context information of the clue instances. However, this work does not aim at classifying sentences or subsentences for subjectivity or polarity like we do here. It focuses on the contextual polarity disambiguation of individual lexicon entries in a text. In both subtasks, we utilize the subjectivity clue lexicon from this work to generate our lexicon-based features. Additionally, we introduce similar syntactic and word context features.

In [4], Breck et al. aims at identifying opinion expressions using a linear-chain conditional random field model. They experiment with the MPQA corpus using lexical, syntactic and dictionary-based features to label each token as *inside* or *outside* an opinion expression. In other words, they mark the boundaries of an opinion expression. Similarly, Choi et al. in [6] utilize the sequential tagging idea in opinion holder identification. Their hybrid approach learns information extraction patterns for extracting the holders, and then applies these patterns as features in their sequential model. Our approach is similar to both [4] and [6] as we exploit sequential tagging at the token level. However, we perform a sentence level classification after token-labeling. Furthermore, instead of conditional random fields we apply SVM^{hmm} which has been shown to perform better in various sequential tagging tasks [1, 19]. Finally, based on the fact that learning tasks, which are highly correlated to the main task, simultaneously, i.e., multi-task learning, increases the performance of the main classification task [5], we learn the opinion expression and the holder labels simultaneously.

3 Approach

We handle sentence level opinionated classification and sentence or subsentence level polarity classification subtasks in two stages. First, our system classifies a sentence as being opinionated or not. Then, we analyze only the opinionated sentences for the polarity classification. The syntactic preprocessing of all documents is done by the TreeTagger[2] POS tagger [17] and the Standford Dependency Parser[3] [13]. We apply a sequential tagging approach in both subtasks.

Formally, the goal of a sequential tagging task is to learn a mapping f from sequences $x \in X$ to discrete outputs $y \in Y$. It is assumed that a training set of input-output pairs $(x_1, y_1), ..., (x_n, y_n) \in X \times Y$ drawn from some unknown probability distribution is available. In both subtasks, we learn labels for individual tokens. Then, we apply a heuristic rule H to propagate the local token level labeling predictions to the sentence or subsentence level classification.

We apply SVM^{hmm} [1, 19] to learn a model from the training samples available in the MPQA corpus. SVM^{hmm} builds on top of SVM^{struct}, which implements a discriminative sequence model that uses the margin maximization approach. Given the training examples $(x_1, y_1), ..., (x_n, y_n)$, SVM^{struct} solves the following optimization problem:

$$min_{w,\zeta} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n} \zeta_i \quad (1)$$

$$s.t. \; \forall (1 \leqslant j \leqslant n), \; \forall y : \quad (2)$$

$$w^T \varphi(x^i, y^j) \geq w^T \varphi(x^j, y) + \Delta(y^i, y) - \zeta_j, \quad (3)$$

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3] http://nlp.stanford.edu/software/lex-parser.shtml

where $\Delta(y^i, y)$ is a loss function calculated as the number of tag differences between y^i and y, c is a tuning parameter for the trade-off between training error and margin. Further details about the algorithm can be found in [19]. The experiments in [15, 19] show that SVM^{struct} performs better than a set of other models like Conditional Random Fields [14] and Averaged Perceptron [7] on a set of sequence labeling tasks.

We differentiate between three types of features in both tasks: *word type* features encoding words as features; *syntactic-lexicon type* features encoding POS as features and the existence of a syntactic dependency between a lexicon instance and the token as features; *lexicon only type* features encoding lexicon look-ups as features.

Most of our features build upon lexicons. Therefore, we first introduce the lexicons in the next subsection. Then, we elaborate on the features used in our formal run configuration in the subsequent subsections.

3.1 Subjectivity Lexicons

Lexicon-based features are built based on two resources: the subjectivity clue lexicons from the previous works [22, 24], hereafter called as the Wilson lexicon, and SentiWordNet [9].

Wilson lexicon consists of three lists of subjectivity clues: *(i)* the prior polarity lexicon, *(ii)* the intensifier lexicon, and *(iii)* the valence shifter lexicon. All parts contain unigram as well as n-gram entries with *POS* and *stemming* attributes. The *POS* attribute indicates the POS of the subjectivity term. The *stemming* attribute indicates whether the look-up should be performed with lemmas or tokens. For instance, the look-up for the lexicon entry (word1=abuse pos1=verb stemmed1=y) should be performed with lemmas and match all the verb instances of the entry like "abused" (verb), "abusing" (verb), but not "abuse" (noun) or "abuses" (noun). Entries of the prior polarity lexicon additionally have the *prior polarity* and *reliability* attributes. *Prior polarity* represents the polarity of an entry out of context with the possible values of *positive*, *negative*, *both* or *neutral*. The *reliability* attribute indicates whether the entry has a subjective usage most of the time *(strongsubj)*, or whether it has only certain subjective usages *(weaksubj)*. The intensifier lexicon contains a list of intensifier words such as "fierce, enormous, more, most". The valence shifter lexicon contains entries which shift the polarity of an existing opinion towards negative or positive including negation words.

Extended Wilson lexicon is a version of the Wilson prior polarity lexicon which we created automatically. We looked up the verbs in the prior polarity lexicon in WordNet to check if they also existed as nouns. Eventually, we added 61 nouns with positive and 192 nouns with negative polarities.

SentiWordNet is a lexical resource which assigns a triplet of numerical scores for positivity *(PosScore)*, negativity *(NegScore)* and objectivity as *(1-(PosScore+NegScore))* to each synset in WordNet. Similar to the Wilson lexicon, SentiWordNet contains unigram as well as n-gram entries with the POS information besides the polarity scores.

3.2 Opinionated Subtask

The opinionated subtask is a sentence level binary classification task, in which each sentence is classified as opinionated or not. We submitted a single configuration in the formal run for this subtask. We utilized a set of linguistic features presented in Table 1 to assign each token a distinct label as: *(i) HOLDER* for belonging to an opinion holder; *(ii) SUBJ* for belonging to a subjective expression; *(iii) NONE* for belonging to none of them. After labeling the tokens as *SUBJ*, *HOLDER* or *NONE*, we applied the following heuristic rule below to compute the sentence level predictions.

$$H(Y_s) = \begin{cases} yes & Y_s \cap \{SUBJ, HOLDER\} \neq \emptyset \\ no & otherwise \end{cases}$$

where Y_s is the label set of all tokens in the sentence s predicted by SVM^{hmm}. A sentence is classified as opinionated if it contains at least one opinionated expression or an opinion holder. Learning opinionated expressions and opinion holders collectively for a sentence level classification can be considered a simple application of multi-task learning. The idea of multi-task learning is to improve the performance of the main classification task by learning a group of highly correlated subtasks simultaneously [5].

Word context, bigram context and POS context features encode information from a window of four tokens. Both *syntactic-lexicon* and *lexicon only type* features exploit the prior polarity information from Wilson's lexicon. *Syntactic-lexicon type* includes four binary features for each token encoding whether the lexicon entry instance is the parent of the token or the child of the token in the dependency parse tree as: *(i) is_modified_by_strongsubj_clue*, *(ii) is_modified_by_weaksubj_clue*, *(iii) modifies_strongsubj_clue*, or *(iv) modifies_weaksubj_clue*. The *lexicon only type* contains six binary features for each token including *is_positive*, *is_negative*, *is_neutral*, and *is_both* for the prior polarity, and *is_strongsubj* and *is_weaksubj* for the prior reliability.

3.3 Polarity Subtask

Similar to the opinionated subtask, we apply SVM^{hmm} to label each token in the polarity subtask as: *positive (POS)*, *negative (NEG)*, or *neutral (NEU)*.

Feature Category	Feature Name	Description	Tasks	Runs
Word	word lemma	lemma of the current token	both	both
	word context	2 tokens to the left and right	both	both
	bigram context	bigram to the left and right	opinionated	both
Syntactic-lexicon	POS	current POS	both	both
	POS context	POS of the 2 tokens to the left and right	both	both
	modified by	modified by weak/strong subj. clue	both	both
	modifies	modifies weak/strong subj. clue	both	both
	negated	if the current token is negated	polarity	both
	preceded by	if the token is preceded by an adj, adv, or an intensifier	both	both
Lexicon only	prior polarity	prior polarity of the current token	both	both
	reliability	reliability of the current token	both	both
	intensifier	if the current token is an intensifier	polarity	both
	valence shifter	if the current token is a valence shifter	polarity	both
	SentiWordNet score	the positivity and negativity score of the term	polarity	first
	extended Wilson nouns	prior polarity of the current token	polarity	first

Table 1. Features of both subtasks

The polarity of a sentence or a subsentence is determined by the following heuristic rule:

$$H(Y_s) = \begin{cases} POS & m(Y_s) > 0 \\ NEG & m(Y_s) < 0 \\ NEU & otherwise \end{cases}$$

$$m(Y_s) = \lambda count(POS, Y_s) - (1 - \lambda) count(NEG, Y_s)$$

where the function $count(l, L)$ counts the occurrences of label l in the label set L. We empirically set λ to 0.5. Since it is assumed that the incoming sentences to be classified in this task are all opinionated, we select only the opinionated sentences from the MPQA corpus during training.

We submitted two run configurations for the polarity subtask. All *word type* features and most of the *syntactic-lexicon* and *lexicon only type* features have the same semantics as described in the previous subsection, as shown in Table 1. Therefore, we will mention only the new features in this subsection. The binary feature *negated* encodes the existence of the following two conditions for each token: *(i)* NEG type dependency relation involving the current token in the dependency tree, or *(ii)* a valence shifter instance of the type negation within a window of four tokens. *Preceded_by* encodes three binary features as *preceded_by_adj*, *preceded_by_adv*, and *preceded_by_intensifier*.

Furthermore, besides the *lexicon only type* features from the first task, we apply binary features for both the valence shifter and the intensifier lexicon look-ups for each token as: *is_ValenceShifter* and *is_intensifier*. Our formal run submission also includes the positivity and the negativity scores from SentiWordNet as double valued features. The same term can appear several times as a member of different synsets with different subjectivity scores in SentiWordNet due to polysemy of words. At this time, we do not perform any word sense disambiguation[4]. *Wilson extended noun* feature consists of two binary features as: *is_positive_WilsonExtendedNoun* or *is_negative_WilsonExtendedNoun*.

4 Experimental Results of Supervised Methods

We have experimented with three corpora (MPQA, NTCIR-6, NTCIR-7) in various settings. We first performed 10 fold cross-validation (CV) experiments for three corpora with the same selection of features using SVM as shown in Table 2. The features include *tfidf*, the number of strong and weak subjectivity clue instances in the current sentence *(#subj)*, the number of strong and weak subjectivity clue instances in the previous sentence and the next sentence *(#context subj)*.

MPQA performs best in CV. Additionally, we observe a significant difference in F-measure when comparing MPQA (F-Measure 0.85) to NTCIR-7 sample data (F-Measure 0.46) in CV. It is, on the one side, the result of overfitting, and on the other side, a problem caused by the low annotation agreement[5] on the NTCIR corpora [18]. Furthermore, in various experiments, we observed that using MPQA as the training corpus always yields better results than using NTCIR-6 as the training corpus when testing on the NTCIR-7 sample collection. Therefore, we report our experi-

[4] SentiWordNet contains synset and sense number information for each term. However, in our experiments we used the polarity information from the first match of the term in the SentiWordNet, we ignore the sense order.

[5] In the NTCIR-7 Test collection, pairwise kappa for the sentence opinionatedness ranges from 0 to 0.45 between three annotators. In the MPQA Corpus, pairwise kappa for the same task ranges from 0.72 to 0.84 between three annotators.

Corpus	Features	P	R	F
MPQA	tf.idf	**0.84**	**0.85**	**0.85**
NTCIR-6	tf.idf	0.51	0.38	0.44
NTCIR-7 Sample	tf.idf	0.54	0.41	0.46
MPQA	tf.idf, #subj, #context subj	**0.85**	**0.86**	**0.85**
NTCIR-6	tf.idf, #subj, #context subj	0.53	0.40	0.45
NTCIR-7 Sample	tf.idf, #subj, #context subj	0.57	0.42	0.47

Table 2. 10-fold cross-validation experiments for the opinionated task on three corpora

ments with MPQA as the training corpus and NTCIR-7 sample data as the test corpus in both subtasks.

Besides sequential SVM^{hmm} models, we experimented with linear SVMs for the opinionated subtask. Table 3 shows a comparison between two models with their best performing feature sets. SVM^{hmm} represents our formal run configuration and the SVMLin[6] represents the linear kernel model with its best performing feature set. Results show that our SVM^{hmm} based experiment setting performs better than the one based on SVMLin on this task increasing both precision and recall.

In another set of experiments, we analysed the contribution of different types of features to the two classification tasks at hand. Table 4 presents our analysis for the opinionated subtask. We started with a configuration which contained only *word type* features without opinion holder identification. Adding *syntactic-lexicon type* features increased both precision and recall. However, adding *lexicon only type* features does not contribute much, while increasing recall. Another major improvement is achieved by including opinion holder identification. Recall is improved by 0.17 at the cost of only 0.03 loss in precision. It is due to the fact that opinion holders occur frequently in the opinionated sentences. Although it is a simple application of multi-task learning, it shows an important characteristic of sentiment analysis: it consists of a set of highly correlated subtasks, which can lead to performance improvement if they are learned simultaneously.

The low precision on the NTCIR data shows that there are much more false positives (objective sentences labeled as opinionated) than false negatives. Most of the false positives contain subjectivity clues used to express facts. For instance, consider the following sentences from the medical and the political domains:

1. *Lung, digestive-tract,* ***blood*** *and skin* ***cancers*** *became increasingly* ***widespread****.*

2. *Iraq has demanded compensation from the U.S. and Britain for the* ***damage*** *caused by their use of* ***depleted*** *uranium shells in air* ***attacks against*** *the country.*

Despite the subjectivity clue instances marked in bold, all with negative prior polarities, both sentences explain facts. With common knowledge we infer that both sentences mention negative developments, but still, they are not opinions. In newspaper articles and in domains like medicine or politics, we see a lot of *negative* or *positive facts*, i.e., *polar facts*, rather than opinions. In polar facts, lexicon entries preserve their prior polarities, but they do not provide reliable evidences for opinions in these domains anymore. We see that the domain knowledge plays a crucial role.

Unlike the strategy given by NTCIR-7 [18], we utilize a different evaluation strategy for the polarity task. We take only the opinionated sentences into account according to the gold standard generated by the lenient method, so that the evaluations of the two subtasks are independent. Table 5 presents our two run submissions for the polarity subtask on the NTCIR-7 sample collection using features presented in Table 1. In this task, our classifier generally tends to label sentences with neutral polarity as opposed to other polarities (especially negative). This is partly due to the fact that the training corpus, MPQA opinionated sample, has a higher proportion of the neutral to polar sentences than the test sample. Low recall of positive sentences and low precision of neutral sentences show that it is much more difficult to differentiate between positive and neutral than between negative and neutral. The result is similar when we apply cross validation in the MPQA corpus. We find less robust evidence of the positive attitude than that of the negative attitude following our approach.

Table 6 presents feature engineering experiments for the polarity subtask. Similar to the opinionated task, we first experimented with *word type* features only. Adding *lexicon only type* features has considerably increased the performance for the classes positive and negative. Obviously, the prior polarity obtained from the lexicons provides more reliable evidence to the classifier than the same information learned directly from the word type features of training samples.

5 Domain Adaptation Experiments

In the previous section, the experimental results show that there is a large discrepancy between the

[6] We used the SVMLin library available at http://people.cs.uchicago.edu/∼vikass/svmlin.html.

Classifier	Features	P	R	F
SVMLin	tf.idf, subj. clue count, SentiWordNet subj. clue count	0.43	0.79	0.56
SVM^{hmm}	word, syntactic-lexicon, lexicon only, holder	**0.45**	**0.84**	**0.59**

Table 3. Comparison of the sequential and linear models

Feature category	P	R	F
Word type features	0.48	0.58	0.53
(+) Syntactic-lexicon type features	**0.50**	0.63	0.56
(+) Lexicon only type features	0.48	0.68	0.56
(+) Holder	0.45	**0.85**	**0.59**

Table 4. Feature selection experiments for the opinionated subtask

MPQA corpus and NTCIR corpora. Since the documents of the two collections cover different topics and the sentiments are generally domain dependent, we try to minimize the discrepancy by using the domain adaptation algorithm Structure Correspondence Learning (SCL) [2, 3], which assumes the data in the target domain is unlabeled.

The goal of SCL is to build a new feature space in which the domain dependent features X_I have a shared representation. Given the labeled data from the source domain and the unlabeled data from both the source and the target domains, a set of pivot features occurring frequently in both domains are selected as the basis of the new feature space. For each selected pivot feature, a linear model is trained on the unlabeled data to predict the occurrence of the pivot in the target domain. Since a trained linear model can be represented as a weight vector, all learned linear models can be represented as a weight matrix, which is further transformed by Single Value Decomposition (SVD) into a projection matrix θ to obtain a low-dimensional dense feature representation. Finally, a discriminative classifier is trained with the new feature representation θX_I instead of the X_I directly for the main classification task. The whole process can also be considered as a way of projecting domain dependent features into a feature space spanned by the pivot features.

The success of SCL relies on the good selection of pivot features, which should frequently occur in both domains and be highly correlated to the main classification task. However, evaluating the selection of pivot features directly through the final classification performance is a quite slow process due to the SVD. We find that the robustness of pivot features can be evaluated by comparing the co-occurrence of pivot candidates and the target classes, because the MPQA corpus, NTCIR-6 and NTCIR-7 sample corpora are annotated. Let PV and OP be the random variables of the occurrences of pivot features and target classes in a certain domain separately, we apply the KL-divergence of the joint distribution $P(PV, OP)$ in both domains.

$$D_{KL}(P\|Q) = \sum_{pv,op} P(pv, op) log \frac{P(pv, op)}{Q(x, op)}$$

where P represents the joint distribution in the source domain and Q is the corresponding one in the target domain. Through our experiments, the best performing pivot features for both tasks are the ones having high conditional probability $P_{position}(OP \mid w)$ in the source domain

$$P_{position}(OP \mid w) = \frac{P(w, OP_{position})}{P(w)}$$

where $P(w, OP_{position})$ is the joint probability of an opinionated expression OP and unigrams and the bigrams w at the position $\{LEFT, MIDDLE, RIGHT\}$ of OP, $P(w)$ is the marginal probability in the MPQA corpus, which is the source domain in our experiments. We build three separate models $\theta_{position}$ for each position in order to predict how likely a pivot feature occurs to the left, right and in the middle of an annotated opinion expression.

Due to the small size of the NTCIR-7 sample data, we define the MPQA corpus as the source domain and the NTCIR-6 corpus as the target domain in our experiments. In both tasks, we train the linear predictors with word and POS features as the non-pivot features. The resulting projection matrices $\theta_{position}$ are used to transform the word and POS features X_w into $\theta_{position} X_w$. The new feature set is then trained again in conjunction with the features in the other two categories with SVM^{hmm}.

Method	P	R	F
supervised	**0.34**	**0.89**	**0.49**
SCL	**0.34**	0.72	0.47

Table 7. SCL experimental results for the opinionated subtask

The experimental results in the two subtasks are compared with our supervised approach in Section 4. As Table 7 shows, we observe a small performance reduction in terms of F-Measure, when we apply SCL in the opinionated task. In the polarity task, in Table 8 we observe that in the transformed feature space, it is

Submission	Positive			Negative			Neutral		
	P	R	F	P	R	F	P	R	F
Run 1	0.80	0.11	0.19	0.79	0.42	0.55	0.07	0.50	0.12
Run 2	0.80	0.11	0.19	0.80	0.43	0.56	0.075	0.50	0.13

Table 5. Comparison of the two polarity run submissions.

Feature category	Positive			Negative			Neutral		
	P	R	F	P	R	F	P	R	F
Word type features	0.75	0.09	0.15	0.79	0.27	0.40	**0.09**	**0.75**	**0.16**
(+) Lexicon only type features	**0.80**	**0.11**	**0.19**	**0.80**	**0.45**	**0.58**	0.07	0.50	0.13

Table 6. Feature selection experiments for the polarity subtask

pivot feature	P_{MPQA}	$P_{NTCIR-6}$	#
urge	1.00	0.53	29
bad	0.97	0.34	35
very	0.80	0.37	76
do not	0.61	0.41	107
should	0.86	0.26	167

Table 9. Example pivot features

more difficult to differentiate polar sentences from the neutral ones. In order to find out the reasons, we compare the conditional probabilities $P(opinionated \mid pv)$ of each pivot feature in the MPQA corpus and the NTCIR-6 corpus. Some frequently occurring example pivot features are given in Table 9, where P_X is the $P(opinionated \mid pv)$ in the corpus X and # denotes the number of the co-occurrences of the pivot feature in both corpora.

We see that there are many pivot features which are good indicators of subjectivity in the MPQA corpus, but strong clues of objectivity in the NTCIR-6 corpus. These pivot features were already used as features in our supervised experiments. For instance, the classifier will learn the occurrence of a feature like "bad" as a strong evidence of subjectivity when trained on the MPQA corpus. Consequently, when the word "bad" occurs in a sentence of NTCIR-6 corpus it will tend to classify the sentence as opinionated. However, only 34% of the sentences containing "bad" are opinionated in the test data. This shows that we already have a strong "overfitting" problem when training on MPQA and testing on NTCIR. SCL makes the situation even worse, because it maps a number of non-pivot features to such domain dependent pivot features. Since the test collection of NTCIR-7 is unlabeled, we cannot judge if a selected pivot feature is domain dependent or not.

6 Conclusions

We proposed a supervised approach based on SVM^{hmm} at the opinionated and the polarity subtasks in the NTCIR-7 MOAT Challenge. It propagates the learned token level labels to the sentence level classification results. In the opinionated task, our system reaches the third place in terms of F-Measure (lenient) and achieves the highest recall 0.91(lenient) among all groups. We also get the second place in the polarity task in terms of F-Measure. Additionally, we present our experiments with structural correspondence learning (SCL) for addressing the domain adaptation problem in sentiment analysis.

Our experiments show that it is a promising approach to learn several correlated subtasks together to achieve a higher performance in the more complex sentiment classification task. As a result, sentiment analysis can be addressed as a multi-task learning problem. The experimental results also show that there is a substantial overfitting problem, when we train our models on the MPQA corpus and test on the NTCIR corpora. Although the low inter-annotator agreement of NTCIR corpora plays an important role in the correct interpretation of our experimental results, we observed that sentiment analysis requires a special type of domain adaptation algorithm, which can solve the problem when the feature vectors x have different $P(y \mid x)$ of the target class y in both domains.

References

[1] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov Support Vector Machines. In *International Conference on Machine Learning (ICML)*, volume 20, page 3, 2003.

[2] J. Blitzer. *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, University of Pennsylvania.

[3] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006.

[4] E. Breck, Y. Choi, and C. Cardie. Identifying expressions of opinion in context. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-2007)*, pages 2683–2688, Hyderabad, India, 2007.

[5] R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

Method	Positive			Negative			Neutral		
	P	R	F	P	R	F	P	R	F
supervised	0.48	**0.16**	**0.24**	**0.78**	**0.49**	**0.60**	**0.27**	0.70	**0.39**
SCL	**1.00**	0.01	0.03	0.75	0.42	0.54	**0.26**	**0.81**	**0.39**

Table 8. SCL experimental results for the polarity subtask

[6] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT 0́5: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[7] M. Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 1–8, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[8] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *CIKM 05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624, Bremen, Germany, 2005.

[9] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, pages 417–422, Genova, Italy, May 22–28, 2006 2006.

[10] A. Esuli and F. Sebastiani. PageRanking WordNet Synsets: An Application to Opinion Mining. In *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*, pages 424–431, Prague, CZ, June 2007. Association for Computational Linguistics.

[11] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Madrid, Spain, 1997. Association for Computational Linguistics.

[12] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, volume 4, pages 1115–1118, 2004.

[13] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

[14] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML-01*, pages 282–289, 2001.

[15] N. Nguyen and Y. Guo. Comparisons of sequence labeling algorithms and extensions. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 681–688, New York, NY, USA, 2007. ACM.

[16] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[17] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of Conference on New Methods in Language Processing*, 1994.

[18] Y. Seki, D. K. Evans, L.-W. Ku, L. Sun, H.-H. Chen, and N. Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In *Proceedings of the 7th NTCIR Workshop, Multilingual Opinion Analysis Task*, 2008.

[19] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. *JOURNAL OF MACHINE LEARNING RESEARCH*, 6(2):1453, 2006.

[20] P. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[21] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 417–424, Philadelphia, Pennsylvania, USA, July 8-10 2002.

[22] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing 2005: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 3406 of *Lecture Notes in Computer Science*, pages 486–497, Mexico City, Mexico, 2005. Springer.

[23] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210, 2005.

[24] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, 2005.

[25] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.