Sentiment Analysis in Chinese

Master's Thesis
Presented to

The Faculty of the Graduate School of Arts and Sciences
Brandeis University
Department of Computer Science
Nianwen Xue, Advisor

In Partial Fulfillment
Of the Requirements for

Master's Degree
By
Lin Pan

February 2012
Abstract

Sentiment Analysis in Chinese

A thesis presented to the Department of Computer Science

Graduate School of Arts and Sciences
Brandeis University
Waltham, Massachusetts

By Lin Pan

Sentiment analysis has been a rapidly growing research area since the advent of Web 2.0 when social networking, blogging, tweeting, web applications, and online shopping, etc., began to gain ever more popularity. The large amount of data from product reviews, blogging posts, tweets and customer feedbacks, etc., makes it necessary to automatically identify and classify sentiments from these sources. This can potentially benefit not only businesses and organizations who need market intelligence but also individuals who are interested in
purchasing/comparing products online. Sentiment analysis is performed on various levels from feature to document level. Supervised, semi-supervised, unsupervised and topic modeling techniques are used towards solving the problems. In this thesis, I explore linguistic features and structures unique to Chinese in a machine-learning context and experiment with document-level sentiment analysis using three Chinese corpora. Results from different feature sets and classifiers are reported in terms of accuracy, which shows the effectiveness of the current approach as compared to traditional machine-learning methods.
# Table of Contents

Introduction......................................................................................................................... 1

1. Sentiment Analysis........................................................................................................... 4
   1.1 Definitions.................................................................................................................. 4
   1.2 The problems.............................................................................................................. 5
   1.3 Supervised approaches............................................................................................... 8
   1.4 Unsupervised approaches.......................................................................................... 12

2. Chinese Sentiment Analysis............................................................................................ 14
   2.1 Corpora...................................................................................................................... 14
   2.2 Word segmentation.................................................................................................... 14
   2.3 Sentiment lexicons.................................................................................................... 15
   2.4 Supervised approaches............................................................................................... 15
   2.5 Rule-based approaches............................................................................................. 16

3. My Approach.................................................................................................................... 20
   3.1 Data.......................................................................................................................... 20
   3.2 Sentiment lexicon....................................................................................................... 20
   3.3 Data pre-processing................................................................................................... 21
   3.4 Classifiers & Evaluation metrics............................................................................... 21
   3.5 Traditional bag of words approach.......................................................................... 22
   3.6 Baseline.................................................................................................................... 23
   3.7 Passive voice............................................................................................................. 27
3.8 Recommend or not? .................................................................................................................. 30
3.9 Dictionary-based features ........................................................................................................ 30
3.10 Referring terms .................................................................................................................... 31
3.11 Discourse relation, punctuation, document-length ...................................................... 31
3.12 Sentiment tracker ................................................................................................................ 33
4. All Results .................................................................................................................................. 39
5. Error Analysis ............................................................................................................................. 42
6. Conclusion .................................................................................................................................... 45
Bibliography ...................................................................................................................................... 47
List of Tables

Table 1: Patterns of POS taggs
Table 2: Accuracies using traditional bag of words
Table 3: Accuracies using baseline new bag of words
Table 4: Dictionary-based features
Table 5: Contrasts, punctuation, doc-length features
Table 6: Degree adverbs classification
Table 7: Computations for opinion phrases
Table 8: Book review corpus results
Table 9: Hotel review corpus results
Table 10: Laptop review corpus results
**List of Figures**

Figure 1: Work flow in supervised sentiment analysis

Figure 2: Sentiment tracker algorithm
**Introduction**

With the rapid development and people's ever growing interest in social networking, blogging, web applications and other information-sharing channels brought about by Web 2.0, more and more time is spent online in people's daily life. This has also contributed to the flourishing online shopping market. However, the problem with online shopping is that people do not get to see the actual products, so we write more and expect more reviews from those who do. Today, large collections of corpora exist with topics ranging from product reviews, movie reviews, to blog posts, tweets, customer feedbacks, etc. *Sentiment Analysis* comes into play when such large amount of data makes it impossible to analyze them manually.

Sentiment analysis has been a widely researched area in English with large amounts of literature on document, sentence, phrase and feature level analyses in different domains (see [2], [3] for a comprehensive reference list). Supervised, semi-supervised, unsupervised and topic-modeling approaches have all been tried out and encouraging results were reported. Automated sentiment analysis could not only benefit businesses and organizations who need market intelligence but also individuals who are interested in purchasing/comparing products. As an example of one of the latest applications of sentiment analysis, *Twitter, inc.* incorporated an advanced tweet-searching function based on sentiment direction, where users can search for positive or negative tweets on a particular topic. A related Internet service
portal in China, Tencent, inc. also recently launched a similar feature on its micro-blogging tool.

In Chinese, research on sentiment analysis is still in the beginning phase ([8], [9], [19], [23], [24]), and most work consists of trying out what have proven to work for English. However, recent years have seen more papers, conferences and talks dedicated to Chinese sentiment analysis ([6], [8], [9], [10], [17], [19], [20], [22], [23], [24], [25]). Chinese Opinion Analysis Evaluation (COAE) is one such conference held by the Chinese Information Processing Society. Up to now, three conferences have been held (in 2008, 2010, and 2011 respectively). The shared tasks include cross-domain (digital, entertainment, finance) sentiment classification on a document level, subjective sentences identification and classification, sentiment word/phrases identification and classification and feature identification ([19]). Out of the 18 participants, most of them focused on the latter three tasks while only 5 presented results on the first task where simple rule-based approaches were taken.

In this thesis, I focus on document-level sentiment analysis in Chinese using three corpora, consisting of book reviews, hotel reviews, and laptop reviews respectively. The goal is to explore linguistic features and structures unique to Chinese that could be efficiently used in a machine-learning based approach. The thesis is organized as follows: Chapter 1 presents the tasks of sentiment analysis and common approaches in general. Chapter 2 presents current research trends in Chinese
sentiment analysis including resources, approaches and corresponding results. Chapter 3 presents my approach including features, classifiers used, and Chapter 4 presents results from different combinations of features. Finally, Chapter 5 presents an error analysis based on the errors made using the best feature set. It is concluded that linguistic features could be highly effective in machine-learning approaches to Chinese sentiment analysis.
1. Sentiment Analysis

1.1 Definitions

Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text. An opinion or sentiment in general is a quintuple of the following ([2], [3]):

(1) $\text{Opinion} = <h_t, e_j, a_{jk}, so_{ijkl}, t_i>$

$h_t$ denotes an opinion holder, i.e., people who expresses subjective opinions on entities, $e_j$ denotes an opinion target, i.e., the entity commented on or reviewed, $a_{jk}$ is an aspect/feature of $e_j$, $so_{ijkl}$ is the sentiment orientation from $h_t$ on $a_{jk}$ of a particular $e_j$, and $t_i$ is the time when the opinion is expressed. For the following example:

(2) I received my Kindle Fire this morning and it is pretty amazing. The size, screen quality, and form factor are excellent. Books and magazines look amazing on the tablet and it checks email and surfs the web quickly.

(from one of the Kindle Fire reviews on Amazon)

$h_t$ would be the reviewer, denoted by $I$, who is commenting on $e_j$, the Kindle Fire, in

---

1 The two terms Sentiment Analysis and Opinion Mining are used interchangeably with the former appearing more in industry and the latter in academia.
terms of its $a_{ij}$, i.e., size, screen, and form. The overall sentiment at the time $t_i$ it was reviewed $s_{\langle ij \rangle}$ is positive.

1.2 The Problems

In English, sentiment analysis is a well-researched area in NLP with a large number of papers, works, etc., and is performed on various levels. Top down, document-level sentiment analysis concerns determining the sentiment orientation of an entire text. Normally, each text is assumed to be completed by a single opinion holder and about a single entity to make the task more tractable. Various machine-learning, unsupervised and semi-supervised approaches exist in the English literature ([1], [5], [11], [13], [16], [21]). The difficulty lies in the fact that there could be mixed opinions in a document, and with the creative nature of natural languages, people may express the same opinion in vastly different ways, sometimes without using any opinion words:

(3) the GMail app is really just a link to Google's GMail website. What on earth were they thinking

(from one of the Blackberry Playbook reviews on Amazon)

Additionally, a text is very likely to contain both objective and subjective sentences at the same time, and we need good selective tools to extract useful information from subjective sentences instead of objective ones. This leads to sentence-level sentiment
analysis.

At sentence level, research has been done on identifying subjective sentences from a mixture of objective and subjective sentences, and then determining the sentiment orientation of the subjective. One may assume that subjective sentence detection can be done via a good sentiment lexicon, i.e., checking if there is overlap between words in the texts and those in the sentiment lexicon, yet the problem is that objective sentences can contain opinion words, as well. For example:

(4) Whether you enjoy exercising or not, you need to do it once a week to keep healthy.

In a given sentiment lexicon, both enjoy and healthy should be listed as positive words. However, in certain contexts, the polarity can be changed (to neutral in this case). One of the approaches towards subjective sentence identification is based on sentence similarity ([27]). The assumption is that subjective sentences are more similar to other subjectives sentences within a topic than to objective sentences. This method requires documents consisting of subjective sentences and those of objective sentences within the same topic as that of the sentence in question. Similarities between sentences are calculated based on shared words/phrases and WordNet synsets to capture the relations among the former. With these two, we average the similarities between the sentence in question and sentences in the documents and assign whichever category that gets a higher score.
The changing of polarity for words in certain contexts reflects the difference between prior polarity and contextual polarity ([15]). Identifying contextual polarities is thus phrase-level analysis. For another example from [15], *unpredictable plot* as in a movie related context is something favorable, but *unpredictable steering* is certainly not in driving.

With document, sentence and phrase-level analyses, we do not know what the opinion holder is expressing opinions on, i.e., the $e_j$ in the quintuple. Furthermore, we do not know what are the features of $e_j$ that are being talked about, i.e., $a_{jk}$. In some applications, for example, a car maker wants to know which particular aspects of their cars do consumers like/dislike, analyses on these two levels are not sufficient. Therefore, we need entity/feature-level analyses. This task involves detecting opinion features and determining the corresponding sentiment orientation. One of the approaches would be to find the terms with highest probabilities on the assumption that features are mentioned more often than other terms in a given review since they are important parts of an opinion target. Another related method is *Association Rule Mining* whose goal is to find correlations between items in large datasets. In general, association rule mining is a two step process. First, we find the most frequent itemsets in a transaction and an itemset is frequent if it meets an user defined *minimum support*. Then, based on *minimum confidence* and the frequent itemsets, association rules are formed. For the purpose of identifying features here, only the first step of the process is needed, that is, finding the most frequent itemsets in a transaction set of nouns/noun phrases. In Hu & Liu ([28]), an itemset is frequent if it appears in more than 1% of the
review sentences. The reason association rule mining is applicable here is that despite the fact that people use different words in a review, when they comment on product features, the words tend to converge ([28]).

In this thesis, I focus on document-level sentiment analysis. In the next section, I will describe common approaches towards this problem in both English and Chinese.

1.3 Supervised Approaches

1.3.1 Classifiers

For supervised document-level classification, three classifiers have shown to be productive and are commonly used: Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM). Steps involved in supervised approaches are data processing, feature extraction, learning/parameter estimation, and decoding as shown in Figure 1 below:

Figure 1: Work flow in supervised sentiment analysis

1.3.2 Naive Bayes
Naive Bayes is a generative model and makes strong independence assumptions. First, we get the model form by using Bayes Rule:

\[ P(label | document) = \frac{P(document | label)P(label)}{P(document)} \]

Each document is represented as a feature vector and by conditional independence assumption between features, we get:

\[ P(label | document) = \prod_{i}^{F} P(f_i | label)P(label) \]

where \( F \) is the total number of features.

For decoding a document, the denominator is the same, so we only need the nominator and pick the label with the highest conditional probability.

In real tasks, features are more or less correlated, yet even so, Naive Bayes tends to perform relatively well ([5]).

**1.3.3 Maximum Entropy**

ME is a discriminative model that directly model \( P(label | document) \), contrary to Naive Bayes where we model the joint probability between the two. Here we require that the model's feature expectation be the same as that of the observed. Among all possible distributions, we select the one with maximum entropy. In other words, we
are committing to the least assumptions possible. The derived model form is as follows:

\[
P(label | document) = \frac{\exp\left(\sum_i^F \lambda_i f(label, document)\right)}{\sum_i \exp\left(\sum_j^L \lambda_j f(L, x)\right)}, \text{ where } F \text{ is the total number of features } f \text{ and } L \text{ total number of } labels.
\]

### 1.3.4 Support Vector Machines (SVM)

SVM is also a discriminative model; Its goal is to find a hyperplane that separates the positive and negative points with the widest margin. In the process, it uses a kernel to map a space of data points in which the data is not linearly separable onto a new space in which it is, with allowances for erroneous classification ([16]).

For text classification problems, SVM is shown to be highly effective: Among all the tests reported in the literature, SVM outperforms Naive Bayes, and ME in most cases ([5], [8], [16], etc.).

### 1.3.2 Features

In a supervised context, feature engineering plays the most important role in obtaining good results. One of the most frequently used methods is a bag of words approach. A bag of words is an unordered collection of words and with this approach to sentiment analysis, each text is treated as such a collection of words. We select a collection of seed words and check their presence in each text. Therefore, this is essentially an
unigram model. The major challenge for this approach is the selection of seed words.

In addition to individual words/unigrams, bigrams, and trigrams are also used in the literature ([5], [21]) as they are useful to capture local dependencies. For example, Pang & Lee ([5]) used unigrams and bigrams on a movie review corpus consisting of 752 positive instances and 1301 negative ones. Unigrams gave the best result with an 83% accuracy on SVM. In [21], NG, et al., also used a movie review corpus consisting of 1,000 positive instances and 1,000 negative ones collected by Pang & Lee\(^2\), and by adding bigrams and trigrams, accuracy went up by nearly 2% from 87.1% to 89.2% over the baseline bag of words using SVM.

Another important finding in Pang & Lee's work ([5]) is that using feature presence is more efficient than feature frequency. By checking unigram presences rather than their frequencies, accuracy went up by 10% on SVM, and 2.3% on Naive Bayes.

In addition to bag of words, part-of-speech is also frequently used ([5], [11], [13]). Pang & Lee ([5]) uses unigrams+POS as features and gets 81.9% accuracy on the movie review corpus using SVM. Yi, et al., ([11]) also use POS patterns as features. Gamon [13] uses POS-trigrams and POS information coupled with semantic relations (e.g. "Verb-Subject-Noun" indicating a nominal subject to a verbal predicate). Intuitively, adjectives/adverbs are strong indicators of subjectivity and opinions, and using adjectives as features is also common practice ([5], [15], [18]). For other features, negation, punctuations and average document length, etc., are also used.

---

\(^2\) Polarity dataset 2.0 available at [http://www.cs.cornell.edu/people/pabo/movie-review-data/]
1.4 Unsupervised Approach  

1.4.1 Point-wise Mutual Information-based  

In information theory, Point-wise mutual information denotes the amount of information that we acquire about the presence of one of the words when we observe the other. In other words, it measures the similarity between words:

\[
PMI(\text{term}_1, \text{term}_2) = \log \left( \frac{P(\text{term}_1 \& \text{term}_2)}{P(\text{term}_1)P(\text{term}_2)} \right)
\]

In [14], Turney uses Point-wise mutual information to compute semantic orientations of particular combinations of phrases, and average all the orientations identified to arrive at the final sentiment orientation.

This approach is based on the assumption that adjectives are important polarity indicators. However, the problem with adjectives in general is that their polarities may change in different contexts. For example, *unpredictable* may imply negativity in the context of *unpredictable steering* for an automobile, yet in a movie review context, *unpredictable plot* is something favorable. To account for this, Turney proposes extracting not only adjectives/adverbs but also a context word if they conform to certain patterns:

Table 1: Patterns of POS tags in [14]
<table>
<thead>
<tr>
<th>First word</th>
<th>Second word</th>
<th>Third word (not extracted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td>NN or NNS</td>
<td>anything</td>
</tr>
<tr>
<td>RB, RBR, or RBS</td>
<td>JJ</td>
<td>Not NN nor NNS</td>
</tr>
<tr>
<td>JJ</td>
<td>JJ</td>
<td>Not NN nor NNS</td>
</tr>
<tr>
<td>NN or NNS</td>
<td>JJ</td>
<td>Not NN nor NNS</td>
</tr>
<tr>
<td>RB RBR or RBS</td>
<td>VB, VBD, VBN, or VBG</td>
<td>anything</td>
</tr>
</tbody>
</table>

It is evident that doing this requires POS-tagging. The next step involves calculating Semantic Orientation according to PMI. The seed word used for positive is excellent, and poor for negative.

\[(9) {SO}(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")\]

SO of a phrase is positive if we get a positive number from this formula and negative if a negative number. The overall orientation of a document is computed by the averaging the SOs of eligible phrases.
2. Chinese Sentiment Analysis

2.1 Corpora

Up to now, there have been two major corpora collected by Tan Songbo at the Institute of Computing Technology, Chinese Academy of Sciences, China, (http://www.searchforum.org.cn/tansongbo/corpus-senti.htm) and by conference organizers of the COAE (Chinese Opinion Analysis Evaluation) (http://www.ir-china.org.cn/coae2008.html#) respectively. Tan's corpus contains 10,000 hotel reviews derived from CTRIP(http://www.ctrip.com/), which is a popular Internet-based travel website in China. Each document is labeled as Positive or Negative according to the number of stars given by consumers; Neutral cases are ignored. The corpus is balanced with duplicates removed. In addition, there are 4,000 laptop reviews from JINGDONG(http://www.360buy.com/), which is a comprehensive online reseller, and 4,000 book reviews, from online book reseller DANGDANG(http://www.dangdang.com/). The COAE corpus contains nearly 40,000 documents on the topics of digital, entertainment, and finance.

2.2 Word Segmentation

As opposed to English, written Chinese does not have word boundaries delimited by spaces; In order to account for individual words, word segmentation is necessary. Besides, some features such as combinations of word/pos requires part-of-speech
tagging. Although these two data pre-processing steps themselves are live research areas, there are many out-of-box tools available for both purposes. One of them is the *Stanford Chinese Word Segmenter*, which is CRF-based, and is implemented in Java. Stanford also provides a Chinese POS-tagger that is based on *Maximum Entropy*. Another important tool is the word segmentation/POS-tagging system by ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System) available at [http://ictclas.org/](http://ictclas.org/). ICTCLAS is based on *Hierarchical Hidden Markov Model* and is implemented in C/C++. Most, if not all papers in the Chinese literature use this system as it is fast and reportedly can achieve up to 98.45% accuracy for word segmentation tasks.

### 2.3 Sentiment Lexicons

Two relatively large-scale sentiment lexicons exist for Chinese sentiment analysis. One of them was compiled by Dong Zhendong, who is the founder of the Chinese HowNet([http://www.keenage.com/html/e_index.html](http://www.keenage.com/html/e_index.html)), and it contains 9,139 opinion words/phrases in Simplified Chinese and 9,142 in English. The other lexicon was compiled by the NLP Lab at National Taiwan University(NTU) ([http://nlg18.csie.ntu.edu.tw:8080/lwku/pub1.html](http://nlg18.csie.ntu.edu.tw:8080/lwku/pub1.html)), and it contains 2,812 positive opinion words/phrases and 8,276 negative ones in both Simplified and Traditional Chinese.

### 2.4 Supervised Approaches
As mentioned in the beginning, sentiment analysis is a growing research area in China, and so far there haven't been much literature covering machine-learning approaches at the document-level. Most of them use rule-based systems. Out of the references I could find, bag of words together with POS and dictionary-based features are used. In [8], Zhang, etc., compares machine-learning approaches with their rule-based system. They experiments with bag of words, word/POS combos and appraisal phrases as features, and finds that word/POS and appraisal phrases give the best result of 83.88% accuracy using SVM classifier and a particular data set.

2.5 Rule-based Approaches

2.5.1 Scoring-based

Classification based on scoring is probably the simplest rule-based approach used in Chinese sentiment analysis. As with the PMI approach, it heavily relies on the sentiment lexicon. Thus in this sense, quality of the lexicon has significant impact on results. Different scoring mechanisms exists: we can simply count the number of positive and negative opinion words in a given document and whichever has a higher count will be the category of the document. If we have a tie, then Negative is assigned, which is an empirical choice (Y. Li, et al., in [19]). Another method is that if zero number of negative opinion words is observed, then the document is classified as Positive; same goes for Negative cases. If both positive and negative opinion words are present, then if the number of positive words is equal to greater than two times that of negative words, then a document is classified as Positive; same goes for
Negative cases. Reason for multiplication and choice of two is to allow for mis-labeled words/phrases in the dictionary, and empirically based, respectively (W. Fei, et al., in [19]). For both of the above methods, they used the sentiment lexicon from HowNet.

In practice, things are too complicated to be predicted by simply counting/comparing opinion words. For example:

\[(10)\] 这本书 工作 不是 很好， 当当 送货 很慢，但是 读了 感觉 还是 非常好。

*This book quality not very good, Dangdang delivery slow, but after reading, feel still very good.*

"This book is not made very well, and Dangdang's delivery was slow, but still a very book after reading it."

Although the sentiment in the first two segments of the sentence is negative, it can be later inverted and possibly the positive aspect far outweighs the negative for the reviewer. By observing this sentence, quality is modified by 不是很好 (*not very well*), delivery is modified by 有点慢 (*a little slow*), and 感觉 (*feeling*) is modified by 非常好 (*very good*). We may want a system that can account for different levels of degree adverbs plus adjectives/adverbs, and based on that, we can detect which aspect the reviewer gives more weight to.

Among other rule-based approaches, Wan proposed an interesting method in [22],
which is to first translate Chinese texts into English using machine translation services (Yahoo Babel Fish and Google Translate), and use a rule-based algorithm on all versions of the text including the Chinese version. From the algorithm, a score is derived for each. The ultimate polarity is determined by the average of each such score. As with other rule-based approaches, the algorithm relies on sentiment lexicons, both in English and in Chinese. It extracts opinion words based on the lexicons and checks if there are negations or intensifiers near the opinion word. If a negation exits, then the score for the opinion word (1 for positive and -2 for negative) is multiplied by -1, and if an intensifier exists, the score is multiplied by 2. The overall score for an entire document is the sum of all these sub-scores from opinion words. When evaluated on a dataset consisting of 886 electronic product reviews from the Chinese it168.com, accuracy reached 85%.

2.5.2 Cascaded Approach

In [8], Zhang, et al., proposes a cascaded rule-based system for document-level sentiment analysis. The overall sentiment orientation of a document is determined by the weighted sum of all the sentences' orientations, which is in turn determined by the opinion words/phrases in the sentence. They use the HowNet sentiment dictionary to detect opinion words and come up with a series of heuristics using dependency trees to determine the polarity of opinion words/phrases. For example, if a word has a VOB (e.g., 是(be) 错误(mistake)) dependency with a child, then the polarity is determined by the child's polarity. At the sentence level, thematic sentences, i.e., sentences that
contain opinion targets, receive more weight than non-thematic sentences, and thematic sentences are defined as the first and last sentences of a document on the assumption that people tend to start or end with topic-related remarks.

Other rule-based approaches for Chinese include using Point-wise Mutual Information, as in [20], where a series of patterns are created, and the overall orientation of a document is determined by the patterned phrases.
3. My Approach

3.1 Data

In this work, I use three product-review related corpora for books, hotels and laptops, respectively. They were all collected by Tan Songbo as mentioned in the previous chapter, and are available online for research purposes at http://www.searchforum.org.cn/tansongbo/corpus-senti.htm. All three corpora are balanced consisting of 2,000 positive reviews and 2,000 negative ones. Duplicate documents were not removed. Ratings were originally expressed with stars and later converted to the categories of positive (4-5 stars) and negative (1-2 stars) and neutral cases were left out. Each data instance is assumed to be from one opinion holder and is about a particular opinion target. The texts come unsegmented. For my experimentation, I split the data into 2,800 reviews for training 1,200 for testing. Upon manual inspection, there are mislabeled documents (6 found, possibly more). For consistency purposes, the data are used as is.

3.2 Sentiment Lexicon

The sentiment lexicon was compiled by Dong Zhendong, who founded the Chinese HowNet (http://www.keenage.com/html/e_index.html). HowNet is comparable to the English Wordnet, and is an online common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in
lexicons of the English and their English equivalents. The lexicon consists of 4,566 positive opinion words/phrases and 4,370 negative ones in simplified Chinese. A point worth noting is that this is a very generic sentiment lexicon and is missing domain-specific words/phrases. For example, words such as 易读 (easy to read), 难读 (hard to read) for book reviews, and 节能 (energy-saving) for laptop reviews are not found in the lexicon. I made some manual changes to the lexicon as it contains duplicate opinion words/phrases in the same category/polarity and also cross-categorically and mis-categorized words/phrases were removed or moved to the right category. In the end, there are a total of 4,420 positive words and 4,300 positive ones in the lexicon.

3.3 Data Pre-processing
As mentioned, the texts are not segmented; Sentence segmentation is first performed, based on punctuation (namely, periods, question marks, exclamation marks, and ellipsis). Then, for word segmentation and POS-tagging, I use the tool developed by Xue ([26]), and it's based on the Maximum Entropy Model.

3.4 Classifiers & Evaluation Metrics
For classifiers, I experiment with Naive Bayes, Maximum Entropy, and Support Vector Machines as they are extensively used for this kind of document classification task. Specifically, for SVM I use SVMlight-6.02 (linear kernel) by Joachims. Both NB and ME are from the Mallet-2.0.7 package by McCallum. Performances from different
corpora and feature combinations are evaluated in terms of *Accuracy*. All features are binary returning either 0 or 1.

### 3.5 Traditional Bag of Words Approach

A *bag of words* is one of most frequently used and efficient approach used in machine-learning based document-level sentiment analysis. In the Chinese literature, Li & Sun ([10]) experimented with a large data set of hotel reviews consisting of 12,000 training instances and 4,000 for testing (both are balanced data sets). They used unigrams, bigrams, and trigrams as features and achieved up to 92% accuracy on the SVM classifier.

To obtain a preliminary result, I use the most frequently used 2,000 words in my training set as a bag of words, plus the question mark as features. Results from NB, ME and SVM\textsuperscript{light} are reported below with the highest accuracy highlighted for each corpus:

<table>
<thead>
<tr>
<th>Corpus Classifier</th>
<th>Book</th>
<th>Hotel</th>
<th>Laptop</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>79.40%</td>
<td>69.25%</td>
<td><strong>62.67%</strong></td>
</tr>
<tr>
<td>ME</td>
<td><strong>82.99%</strong></td>
<td>72.5%</td>
<td>62.33%</td>
</tr>
<tr>
<td>SVM\textsuperscript{light}</td>
<td>80.58%</td>
<td><strong>73.42%</strong></td>
<td>61.58%</td>
</tr>
</tbody>
</table>
3.6 Baseline

3.6.1 New Bag of Words Extraction

One of the problem with the bag of words above is that although the extracted words can be domain-specific as they are from the training data, there are still quite a number of uninformative features, such as the word 想 (think), which possibly preludes an opinion but not part of it. Furthermore, more features are to be added based on the HowNet lexicon and as previously mentioned, the lexicon is very generic and missing domain-specific words/phrases. In order to have a bag of words that are domain-specific, informative and at the same time can complement the generic HowNet lexicon, I experiment with a new bag of words derived from three particular linguistic structures: Chinese parallelism sentences, causatives, and words immediately following degree adverbs.

3.6.2 Parallelism Sentences

A parallelism sentence in Chinese is defined as a sequence of three or more semantically or structurally similar phrases/sentences. The benefit of using parallelism is to promote clarity, vividness and compactness in expressing emotions. To illustrate the points, let's see an example:

(11) 这 本书 价格/NN 便宜/JJ, 做工/NN 精良/JJ, 意义/NN
深刻/JJ。

This book price/NN cheap/JJ, craftsmanship/NN good/JJ, meaning/NN
This book is reasonably priced, well made, and profound in meaning.

Starting from the third character of the Chinese sentence (price/price), each phrase of the sequence has the same structure of a noun (NN) followed by an adjective (JJ), separated by a comma with each phrase describing the attribute that the nouns have. These nouns and adjectives in the structure (highlighted in bold) are extracted. Using parallelism is probably the most compact yet clear way of expressing the idea conveyed by the example sentence. For the same meaning, I could have said:

(12) 这本书 价格/NN 便宜/JJ，做的/VB 也不错/JJ，还有/VB 深刻的/JJ 意义/NN。

This book price cheap, made well, besides, has profound meaning.

For a native Chinese speaker, these two sentences mean the same. However, the first one is more organized and succinct while the second is more likely to appear in colloquial contexts. In practice, the structural parallelism can also be extended to a sentence level, i.e., three or more sentences that have the same structure, but there are few instances of this in the training set, so during feature extraction, I only included inner-sentence parallelism cases like the first example above and extracted the words in pattern. Note that the parallelism is not limited to NN followed by JJ, but also any
other combinations that recur three or more times in a sentence and are separated by commas or semi-colons.

3.6.3 Causative structures

In Chinese, causatives are expressed by 使, 令, 叫 and 让 both denoting make in the English sense. For example:

(13) a. 这本书让我体会到了人生的哲理。

This book made me realize life’s maxims.

"This book made me realize the maxims of life."

b. 于丹的书令我感动！

Yu Dan's book made me move!

"Yu Dan's book moved me!"

There are subtle differences between the three characters (mostly stylistically), but due to the scope of this work, they are not discussed. Usually 使, 令, 叫 and 让 are followed by a noun denoting the affected and a verb or adjective denoting the action or resulting state. As can be seen from the above examples, the words following these characters can potentially be opinion words. For this step, it extracts the two words following these four key words and this is performed on all the training instances. In the above two examples, extracted words are highlighted in bold.
3.6.4 Word after Degree Adverbs

Degree adverbs frequently accompanies adjectives, which are good indicators of subjectivity and opinion([5], [14], [18]). Chinese has a rich set of degree adverbs, as shown in (14), where for a single meaning, there can be multiple words; This potentially makes sentiment classification more difficult since we need to account for the subtle degree differences to arrive at the right polarity. This feature extraction step is very straightforward: It extracts the word immediately after degree adverbs, and the degree adverbs are lists below³:

(14) Chinese_degree_adverbs = [真(really), 最, 最为(most), 极, 极为, 极其, 极端, 至为, 顶, 过, 过于, 太, 过分, 分外, 万分(extremely/too), 更, 更为, 更加, 更其, 越, 越发, 越加, 倍加, 愈, 愈发, 愈加, 愈为, 愈益, 益发(more), 格外, 很, 挺(very), 老(often), 非常, 特别, 相当, 十分, 甚为, 颇, 颇为, 异常, 深为, 满, 蛮, 够, 多么, 特, 大为, 何等(very), 有点, 有些, 稍, 稍稍, 稍微, 稍许, 稍为, 略, 略略, 略微, 略为, 些微, 多少(a bit), 较, 较为, 较比, 较为(comparatively), 还, 不太, 不大, 不很, 不甚(not very)]

3.6.5 New Bag-of-Words Results

The words extracted from the Chinese causatives, parallelism structures and words following degree adverbs from the training data thus form a new bag of words. The following table shows results using NB, ME and SVMlight respectively:

---
³ If a Chinese word is not translated, it means that it has the same meaning as its following words until a translation is provided.
Table 3: Accuracies using baseline new bag of words:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Book</th>
<th>Hotel</th>
<th>Laptop</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>86.5%</td>
<td>82.75%</td>
<td>89.67%</td>
</tr>
<tr>
<td>ME</td>
<td>90.83%</td>
<td>82.42%</td>
<td>87.33%</td>
</tr>
<tr>
<td>SVM&lt;sub&gt;light&lt;/sub&gt;</td>
<td>88.41%</td>
<td>81.83%</td>
<td>86.75%</td>
</tr>
</tbody>
</table>

The result shows that the new bag of words extracted from these Chinese linguistic structures substantially improves performance over the traditional bag of word approach for all three corpora and classifiers. For the book review corpus, accuracy improves by 7.1% on NB, 7.84% on ME, and 7.83% on SVM<sub>light</sub>. For the hotel review corpus, accuracy goes up by 13.5% on NB, 9.92% on ME, and 8.41% on SVM<sub>light</sub>. Same trend goes on for the laptop review corpus, with up to 27% accuracy increase on NB, 25% on ME, and 25.17% on SVM<sub>light</sub>.

### 3.7 Passive Voice

In English, *passive voice* is normally formed with the *be + past participle* construction:

*(15) Mary was given a book by John.*
(16) It is said that he left.4

In Chinese, it is denoted by the word 被(be) alone, since there are no morphological inflections in Chinese.

(17) 我被故事中的主人公给感动了。

I was story inside main character moved.

"I was moved by the main character in the story."

One may assume that usage of passive voice is the same in the two languages. However, neither one of the English examples can be directly translated into Chinese using 被.

(18) *玛丽被约翰给了本书。

Mary was John given a book.

"Mary was given a book by John."

(19) *被说他走了。

It is said he left.

"It is said that he left."

The first Chinese sentence must be in the active voice:

4 Both examples are from Wikipedia: http://en.wikipedia.org/wiki/English_passive_voice
(20) 约翰给了玛丽一本书。

"John gave Mary a book."

In the second sentence, 被 should be replaced by 据 (according to).

It seems that passive voice can be more freely used in English compared to Chinese, and 被 is more associated with the benefactive/malefactive. In other words, 被 seems to appear more in negative contexts as opposed to positive ones.

(21) 看了这本书有种被骗的感觉。

Read this book has kind of be cheated feeling.

"After reading this book, I felt cheated."

To capture this difference, I made a list of verbs that when used with 被 express positive opinions and the feature function will return 1:

(22) passive_positive = [教育(educate), 震惊(shocked), 感动(move), 打动(move), 折服(impress), 吸引(attract), 表扬(praise), 批准(approve)]

Other than these verbs, it is assumed that passive voice denotes negative opinion and the feature function returns 0.
3.8 Recommended or Not?

In product reviews, it is commonplace that people would like to share and recommend the products they are satisfied with and warn against those that they are not, or recommend other products instead.

(23) 这本书 非常 好，强烈 推荐！！

This book very good, highly recommend!!

"This book is very good, highly recommend it!!"

This feature checks if negation is present or not before 推荐；return 0 if present and 1 otherwise.

3.9 Dictionary-based Features

Features presented in this section are entirely based on the sentiment lexicon from HowNet. More specifically, two features check if positive and negative words are present in a given document; One feature compare the number of polarity words; And the other feature checks the polarity of the last opinion word in a document. They are summarized in the following table:

Table 4: Dictionary-based features

<table>
<thead>
<tr>
<th>Feature</th>
<th>What it checks</th>
<th>Return_val</th>
<th>Return_val</th>
</tr>
</thead>
<tbody>
<tr>
<td>nos_presence</td>
<td>positive words presence</td>
<td>1 if present</td>
<td>0 if not</td>
</tr>
</tbody>
</table>

30
<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>neg_presence</td>
<td>negative words presence</td>
<td>1 if present</td>
</tr>
<tr>
<td>compare_num</td>
<td>which polar has greater number</td>
<td>1 if more pos</td>
</tr>
<tr>
<td>last_polar</td>
<td>polar of last opinion word</td>
<td>1 if pos</td>
</tr>
</tbody>
</table>

**3.10 Referring Terms**

This is a highly domain-specific feature for the book review corpus. In Chinese, a book can be referred to in a number of ways: 著作 (book) is the most generic term. In addition to 著作, the following list of words can also be used to refer to a book:

\[(24) \text{book\_referring\_terms} = \{作品(\text{work}), 大作, 巨作, 巨著, 著作, 名作, 名著(\text{masterpiece})\}\]

The first Chinese word in the list is comparable to the English work when used to refer to a book. From the second word on, all the words have a similar meaning to the English masterpiece. Again, subtle differences between these Chinese words are not discussed here. This feature checks if a reviewer uses one of these referring terms when mentioning a particular book.

**3.11 Discourse relation, Punctuation & Document length**

This set of features captures information about discourse relation, punctuation, and document length. Clearly, contrasts in discourse relations are important in expressing opinions in that they invert the polarity of these opinions, and the contrast feature

---

31
checks if a document contains two or more contrast terms, which are listed below:

(25) **Contrasts** = [但，但是，可是，然而，不过，却，偏偏，只是，至于，不料，岂知](however/but/yet)]

The **punctuations feature** checks if multiple punctuations exist in a document, such as two or more ellipses⁵ (marked by six or more Chinese periods), three or more question marks or exclamation marks. Empirically, people use multiple punctuations to emphasize certain emotions. For example:

(26) 我
不
知道
该怎么办了。.....

I  not know  what to do.....

"I don't know what to do....."

(27) 这 本书 也  叫  大作吗？？

*This book  as well called masterpiece??*

"You call this book a masterpiece??"

Finally, the document-length feature checks if the length of a document exceeds a certain threshold. Here I use 80 as the threshold, which is the average length of the negative reviews in the book review corpus. The following table summarizes this set of features:

---

⁵ Periods in Chinese are marked by "。" instead of the English ".". Ellipsis in Chinese is usually marked by six periods "。。。。。。", yet in casual writing, people often use three periods as one ellipsis mark. Thus in this thesis, the latter expression is assumed.
Table 5: Contrast, Punctuation, and Doc-length features

<table>
<thead>
<tr>
<th>Feature</th>
<th>What it does</th>
<th>Return</th>
<th>Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrasts</td>
<td>if doc contains 2 or more</td>
<td>1 if true</td>
<td>0 otherwise</td>
</tr>
<tr>
<td>Punctuations</td>
<td>if doc contains multiple punctuation</td>
<td>1 if true</td>
<td>0 otherwise</td>
</tr>
<tr>
<td>Doc-length</td>
<td>if doc contains 80 or more words</td>
<td>1 if true</td>
<td>0 otherwise</td>
</tr>
</tbody>
</table>

3.12 Sentiment Tracker

In section 2.5.1, I mentioned that a rule-based system needs finer-grained classification schemes than simply counting the number of positive and negative words and then comparing those two numbers as documents with more positive words than negative ones may express an overall negative sentiment and vice versa. Furthermore, in a mixture of positive and negative words, we need to give higher weights not only to subjective sentences than objective ones but also opinion words with higher degrees as illustrated in example (10), repeated as (28) below:

(28)这本书做工不是很好，当当送货很慢，但是读了感觉还是非常好。

"This book is not made very well, and Dangdang's delivery was slow, but still a very good book after reading it."

In order to capture the degree difference in 不是很好 (not very good), 慢 (slow), and 非常好 (very good), and come down to 非常好 (very good) in the end, we first
categorize the degree adverbs into four degree levels as shown in Table 6, and then a scoring system is created based on opinion words, negation words and degree adverbs. These three elements form an opinion phrase.

Table 6: Degree adverbs classification

<table>
<thead>
<tr>
<th>Extreme</th>
<th>最, 最为(most), 极, 极为, 极其, 极端, 太, 至为, 顶, 过, 过于, 过分, 分外, 万分, 超, 超级(extremely)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>真(really), 更, 更为, 更加, 更其, 越, 越发, 倍加, 愈加, 愈, 愈发, 愈为, 愈益, 越加(more), 格外, 益发, 很, 拥, 老, 非常, 特别, 相当, 十分, 甚为, 颇, 颇为, 异常, 深为, 满, 蛮, 够, 多么, 特, 大为, 何等(very)</td>
</tr>
<tr>
<td>Medium</td>
<td>较, 比较, 较比, 较为(comparatively), 还, 不大, 不太, 不很, 不甚(to some extent)</td>
</tr>
<tr>
<td>Low</td>
<td>有点, 有些, 稍, 稍稍, 稍微, 稍为, 稍许, 略, 略略, 略微, 略为, 些微, 多少(a little bit more or less)</td>
</tr>
</tbody>
</table>

With this table, we can start computing polarity scores for opinion phrases. In specific, we extract opinion words from a document based on the sentiment dictionary; how and which opinion words are extracted will be talked about later. Then we extract the
two words before these opinion words to form an opinion phrase. The next step is to score the opinion phrases according to the table below:

Table 7: Computations for opinion phrases

<table>
<thead>
<tr>
<th>Opinion Phrase</th>
<th>Computation</th>
<th>Example</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;+&quot;+O</td>
<td>S(O)</td>
<td>book quality <strong>good</strong></td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>book quality <strong>poor</strong></td>
<td>-0.8</td>
</tr>
<tr>
<td>&quot;+NA+O</td>
<td>S(NA)*S(O)</td>
<td>book quality <strong>not good</strong></td>
<td>-0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>book quality <strong>not poor</strong></td>
<td>0.64</td>
</tr>
<tr>
<td>NA+NA+O</td>
<td>S(NA)*S(NA)*S(O)</td>
<td>book quality <strong>not not good</strong></td>
<td>0.512</td>
</tr>
<tr>
<td>&quot;+DA+O</td>
<td>S(DA)*(1-S(O)+S(O)</td>
<td>book quality <strong>very good</strong></td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>S(DA)*(-1-S(O))+S(O)</td>
<td>book quality <strong>very poor</strong></td>
<td>-0.94</td>
</tr>
<tr>
<td>NA+DA+O</td>
<td>0.5+0.2*S(DA)</td>
<td>book quality <strong>not very good</strong></td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>-0.5+(-0.2)*S(DA)</td>
<td>book quality <strong>not very poor</strong></td>
<td>-0.64</td>
</tr>
<tr>
<td>DA+NA+O</td>
<td>-0.64+2<em>0.2</em>(-0.8)*S(DA)</td>
<td>book quality <strong>very not good</strong></td>
<td>-0.864</td>
</tr>
<tr>
<td></td>
<td>0.64+2*(-0.2)*(-0.8)*S(DA)</td>
<td>book quality <strong>very not poor</strong></td>
<td>0.864</td>
</tr>
</tbody>
</table>

In this table, O denotes opinion words, NA negation words and DA degree adverbs. The empty strings "+" indicate that the word is not in any one of the three categories above. S is the scoring function. Scores for the four degrees **extreme, high, medium**, **not very good, not very poor, very not good, very not poor**

---

6 Due to space limit, Chinese texts are not provided here and the examples are gloss translations. The double negation is not allowed in the form of "not not" in English, but it is how it is used in Chinese.
and low are 0.9, 0.7, 0.5 and -0.5 respectively. Positive words receive a score of 0.8 and negative ones -0.8. Score for negation is -0.8 rather than -1 because 不好 (not good) is not the same as 差 (poor).

With these scoring mechanisms, this feature is an algorithm that takes a cascaded approach building from opinion phrases, to sentences and the polarity of the entire document is determined by the strongest opinion. In the process, subjective sentences receive more weight than objective sentences. Pseudo-code for this algorithm is presented in Figure 2:

Figure 2: Sentiment Tracker algorithm

Targeted_opinion = []
Non_targeted_opinion = []

1. look for opinion targets in the first sentence
2. if found, look for opinion words in a window of three segments. Extract opinion phrases (opinion word plus two words before it) and compute polarity scores for them according to above table. Select the strongest one and append it to targeted_opinion.
3. if not found, look for opinion words in the sentence and do the same computation and appending as above.
4. look for opinion targets in subsequent sentences,
5. if no opinion targets are found:
   Look for opinion words in subsequent sentences and append them to
non_targeted_opinion list

6. if found:

   for each opinion target:

      do step 2

      do step 3 for last sentence

7. find highest score in targeted_opinion T and non_targeted_opinion N

8. return the greater of T and 0.96 * N

In step 2, each segment is defined as any sequence of words that is delimited by any punctuation mark. Finding opinion targets themselves is a live research area as mentioned before, and the method used here for finding opinion targets is oversimplified. Opinion words are found it if one of the two conditions is satisfied: (1) if a word is in list (29) for the book review corpus. (2) if we find the pattern of \{这个(this)|这个(this)|本(this)|部(collection)|套(collection)|书(book)\} for the book review corpus, and the pattern of \{这个(this)|家(this)|此(this)|本(this)|酒店(hotel)|饭店(hotel)|旅馆(hotel/motel)|旅馆(hotel/motel)|宾馆(hotel)\} for the hotel review corpus, and \{这个(this)|这台(this)|这台(this)|电脑(computer)|笔记本(laptop)|机器(machine)|子(machine)|本子(laptop)|系统(system)\}, then an opinion target is found.

(29) \textit{opinion_target} = [作者(author), 里面(inside), 作品, 著作, 名作, 大作, 巨著,
Here I'm assuming that sentences with opinion targets are subjective sentences and those without opinion targets are objective ones. This algorithm gives less weight to objective sentences by multiplying 0.96 to the strongest opinion phrase in the last step of the algorithm. This feature returns 1 if the greater of T and 0.96*N is a positive number and 0 if negative. As a result, this feature can also be used as a rule-based system, and for the test set of the book review corpus, it gives 63% accuracy.
4. All Results

This section presents results from different combinations of features mentioned above. In the tables below, accuracies from different classifiers on all three corpora are reported. Rec is the Recommend or not feature, Passive the Passive voice feature, Dict the set of dictionary related features, Ref the Referring term feature, Punc the Long punctuation feature, and Len the Document length feature. Note that for the book review corpus, there is a domain-specific feature, namely, the Referring terms feature. For the majority of the tests, more features means better results, yet it's not always the case. In the hotel review corpus, NB gets the best results from the new bag of words alone and the sentiment tracker feature hurts accuracy on all three classifiers for the laptop review corpus.

The best result is up to 92.25% accuracy for the book review corpus using all features on ME, 84.08% for the hotel review corpus using all features on ME, and 90.58% for the laptop review corpus using all features except for the sentiment tracker feature on NB.

Table 8: Book review corpus results

<table>
<thead>
<tr>
<th>Feature set</th>
<th>NB</th>
<th>ME</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 2,000 words + &quot;!&quot;</td>
<td>79.83%</td>
<td>83.33%</td>
<td>80.58%</td>
</tr>
</tbody>
</table>
### Table 9: Hotel review corpus results

<table>
<thead>
<tr>
<th>Feature set</th>
<th>NB</th>
<th>ME</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 2,000 words + &quot;!&quot;</td>
<td>69.25%</td>
<td>72.5%</td>
<td>73.42%</td>
</tr>
<tr>
<td>New bag of words/Baseline</td>
<td><strong>82.75%</strong></td>
<td>82.42%</td>
<td>81.83%</td>
</tr>
<tr>
<td>Baseline+Rec+Passive</td>
<td>82.58%</td>
<td>82.58%</td>
<td>81.67%</td>
</tr>
<tr>
<td>Baseline+Rec+Passive+Dict</td>
<td>81.75%</td>
<td>83.42%</td>
<td>82.67%</td>
</tr>
<tr>
<td>Baseline+Rec+Passive+Dict+Ref+Contrast+Punc+Len</td>
<td>81.67%</td>
<td>83.25%</td>
<td>83.17%</td>
</tr>
<tr>
<td>All features</td>
<td>82.17%</td>
<td><strong>84.08%</strong></td>
<td><strong>83.75%</strong></td>
</tr>
</tbody>
</table>

### Table 10: Laptop review corpus results

<table>
<thead>
<tr>
<th>Feature set</th>
<th>NB</th>
<th>ME</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 2,000 words + &quot;!&quot;</td>
<td>62.67%</td>
<td>62.33%</td>
<td>61.58%</td>
</tr>
<tr>
<td>Feature Set</td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>----------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>New bag of words/Baseline</td>
<td>89.67%</td>
<td>87.33%</td>
<td>86.75%</td>
</tr>
<tr>
<td>Baseline+Rec+Passive</td>
<td>89.5%</td>
<td>87.67%</td>
<td>86.58%</td>
</tr>
<tr>
<td>Baseline+Rec+Passive+Dict</td>
<td>89.58%</td>
<td>87.67%</td>
<td>86.83%</td>
</tr>
<tr>
<td>Baseline+Rec+Passive+Dict+Contrast+Punc+Len</td>
<td>90.58%</td>
<td>88.67%</td>
<td>87.83%</td>
</tr>
<tr>
<td>All features</td>
<td>90.5%</td>
<td>88.08%</td>
<td>87.5%</td>
</tr>
</tbody>
</table>
5. Error Analysis

By observing errors made by the classifiers, some of them are very difficult for machine-learning systems to extract useful and consistent features from, and for the rule-based ones to capture subjective contents from. Take one of the mis-classified cases:

(30) 这本书非常适合 7 岁以下的儿童阅读，如果你 7 岁以上，还是看看别的书吧。

This book very suitable 7 years old below children read, if you 7 years old above, read other books.

"This book is very suitable for children under 7 years old. If you are older than that, you better look elsewhere."

The book being reviewed (This book) is intended to people of all ages. With that, it is evident that this utterance expresses a negative sentiment and it involves the use of irony. With the rule-based algorithm, this will be classified as positive since 适合 (suitable) is extracted as a positive word modifying the opinion target 这本书 (this book) and other than that, there is no opinion word, hence positive overall. Without a way to get onto the pragmatic level, rhetoric poses great challenges for both supervised and rule-based approaches.
Among other errors, comparatives and superlatives are difficult to capture, in that in some cases it involves correctly identify the opinion target. For example:

(31) 《傅佩荣 <论语> 心得》比《于丹<论语>》深刻，如果还是小学，初中水平，建议你看看于丹的，高中以上，建议你看傅佩荣先生著作系列。

"<Fu Peirong <Analects of Confucius> Reflections> compared <Yu Dan<Analects of Confucius>Reflections> more profound, if still grade school, junior high school level, recommend you read Yu Dan's, high school above, recommend you read Fu Peirong masterpiece collections.

"Compared with <Yu Dan<Analects of Confucius>Reflections>, <Fu Peirong<Analects of Confucius>Reflections> is more profound in meaning, if you are on a grade school or junior high school level, then I recommend you look into Yu Dan's book. If you are above high school level, I recommend you read Fu Peirong's masterpiece works."

This is a negative review for the book <Yu Dan<Analects of Confucius>Reflections>. For this example, it is not immediately clear which one of the two books is the one being reviewed on, that is, the opinion target. In fact, without prior knowledge or more contextual clues, either one could be. This presents much difficulty for both rule-based and machine-learning based approaches. For rule-based,
it will capture 深刻 (profound) and it's polarity, which is positive. Other than that, there is no opinion words, so the overall orientation as predicted by the rule-based system would be positive. As for features, we see 著作 (masterpiece), which is used as a feature for the Referring Terms feature above, and 深刻 (profound), as a positive word. Both are clues for a positive review. Besides, another problem with feature extraction for comparatives and superlatives is that there are relatively few instances of them, so it may not be an informative feature. Therefore, without being able to correctly identify the opinion target, in a comparative context, it's easy to assume the wrong polarity.
6. Conclusion

In this thesis, I experimented with document-level sentiment analysis for three product review related corpora. A new set of linguistics-oriented features were used in a machine-learning based approach and they proved to be efficient with accuracies of up to 92.25% for the book review corpus, 84.08% for the hotel review corpus and 90.58% for the laptop review corpus, respectively. The new bag of words extracted from the three Chinese linguistic structures form a much stronger baseline against the traditional bag of words approach. This new bag of words are domain-specific as they are extracted from the training corpus and it can complement the generic HowNet or other sentiment dictionaries. In addition, it is easy to migrate this method to other domains.

In a product review-related context, it is common that people tend to recommend products they are satisfied with and warn against those that they are not. A feature was used to capture that information. In Chinese, it seems that the passive voice is used somewhat differently from how it is in English in that it's closer to the benefactive/malefactive in the majority of cases, and extracting features from voice information proved to be useful together with the Recommend feature. A set of dictionary-based features also contributed to an accuracy increase. A list of book referring terms were used specifically for the book review corpus. Discourse relation (contrast), punctuations, and document-length were also captured as features. The last
feature, *sentiment tracker* contributed to accuracy increase for two corpora while hurting that of the laptop review corpus.

As argued in the previous chapter, comparatives and superlatives present great difficulty for both rule-based and machine-learning based approaches. Due to their relatively low frequency, a rule-based system may be in a better position to account for it, and to do so requires a more accurate opinion target identification system. What seems much more difficult is to get onto the pragmatic level in order to account for rhetorics, such as irony. Correctly learning/predicting cases like these requires more complicated systems that is left for future work.
Bibliography

Language Processing, 2005.