Theme Based Text Sentiment Analysis

Harbin Institute of Technology (Shenzhen)

Intelligent Computation Research Center

17S051048 冷 佳

17S151549 彦虹羽

17S151539 龚晓畅

Date: 12/09/2017

Contents

[1. Introduction 1](#_Toc500710908)

[2. Data analysis 1](#_Toc500710909)

[3. Implement 3](#_Toc500710910)

[3.1 General frame 3](#_Toc500710911)

[3.2 Data processing 4](#_Toc500710912)

[3.3 Segmenting the words 6](#_Toc500710913)

[3.4 Identification of Themes and Sentiment Words 6](#_Toc500710914)

[3.5 The Matching of Themes and Sentiment Words and Polarity tagging 14](#_Toc500710915)

[4. Results 17](#_Toc500710916)

[5. Conclusion 19](#_Toc500710917)

[6. Task Assignment 20](#_Toc500710918)

# 1. Introduction

In recent years, the text sentiment analysis plays an increasingly important role in network marketing, enterprise public opinion monitoring and government public opinion monitoring. The text sentiment analysis based on the theme becomes the focus of hot spots due to the advantage of subject model in the field of text mining. Its main task is to improve the performance of text sentiment analysis, by mining the theme in users’ comments, as well as the emotional preference of these themes.

This competition is based on e-commerce review data. We need to analyze the users’ preference for these themes according to the statement of subject characteristics and emotional information. Finally, we should output the pairs of themes, sentiment words and sentiment polarity.

# 2. Data analysis

The dataset of this competition is the desensitized e-commerce review data. The structure of training data is shown as follows. The number of themes and sentiment words can be more than one, and each theme must be paired with corresponding sentiment word and sentiment polarity. The themes can be null, but sentiment words can’t.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Row\_id  (String) | Content  (String) | Theme  (String) | Sentiment word  (String) | Sentiment anls  (String) |
| 15 | 表带带着痒，不知道怎么回事 | 表带 | 痒 | -1 |
| …… | …… | …… | …… | …… |

Figure 2-1 The structure of training data

The contents of training data is informal and colloquial, which results in high noise in data. By analyzing training data, we find that some extracted themes and sentiment words are incorrect obviously. The example shown as follows is incorrect. The extracted sentiment word is “不能便宜货”, which should be changed into “不能用”. The errors similar to above example make a challenge to train models.



Figure 2-2 The incorrect example

The structure of test data is shown as follows. It only have ids and contents. Hence, we should extract the pairs of themes, sentiment words and sentiment polarity correctly, which is the goal of this competition.

Table 2-1 The structure of test data

|  |  |
| --- | --- |
| row\_id  (String) | Content  (String) |
| 7 | 此产品太差劲了，没有任何内包装，直接放在盒子里就过来了。铅笔也是打开过多次，已经很久了的样子。必须差评！ |
| …… | …… |

# 3. Implement

## 3.1 General frame

By analyzing training data, we decide to use rule based and CRF based ways to get the results. The preprocessing, words segment and matching pairs are similar between two ways. There are differences in identification of themes and sentiment words which will be explained in following content.

The basic frame of our ways are shown as follows.

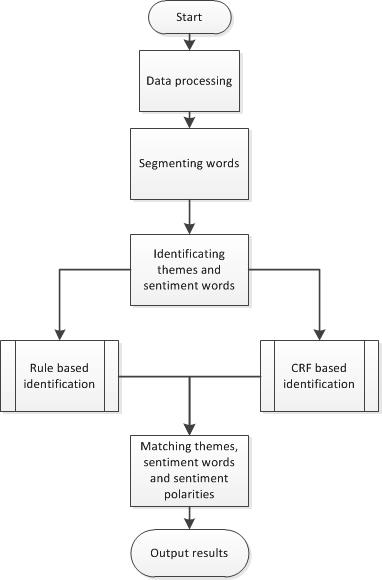


Figure 3‑1 The basic frame of our ways

## **3.2 Data processing**

For the electricity comments, we process the raw texts via the following steps:

1. Remove all useless information shown in the following Table 3-1
2. Using regular expression to replace simple punctuation or continuous punctuation that can split the sentences into sub-sentences with one comma. We think that punctuation contains “?、？、!、！,、，、.、。”.
3. Lower all upper cases.
4. Convert traditional Chinese into simple Chinese with the tool opencc.
5. Remove all numbers.

Table 3-1 The removing information

|  |  |
| --- | --- |
| Category | Instance |
| Expressions | &hellip;&hellip;&hellip; |
| Additional review | [追评] |
| URLs | http://www.amazon.cn |
| @tags | @小桀 |
| Titles | 【PSV国行游戏】 |
| Newline symbol | </br> |
| The hashtag # # | #疾速性能# |

After data processing, we can get some empty columns, and then we move these instances into a null\_data.csv.

## 3.3 Segmenting the words

Before segmenting the words, we split the sentence into some sub-sentences in order that matching the themes and sentiment words better. We split the sentence according the comma get in last step.

The jieba tool can identify some new words and we can also add Custom dictionary to segment the word better. Therefore, we use jieba tool to segment the words.

The comments are about appliances, clothes and phone etc. There are many corpus words so that we collect the theme and sentiment words from training data whose frequency is more than 50. Then we add these words into Custom dictionary.

After segmenting the words, there many stop words such as “啊 的 呢 很 太 再” etc. which don’t have actual functions. Thus, we remove these stop words as following step.

## 3.4 Identification of Themes and Sentiment Words

**3.4.1 CRF**

1. **Description**

Conditional random fields (CRFs) are a class of statistical modeling method often applied in pattern recognition and machine learning and used for structured prediction. In this experiment, we use CRF to tag theme words and sentiment words.

First, we extract some features about every word in a sentence from the training data.

In a CRF, each feature function is a function that takes in as input:

* a sentence s
* the position i of a word in the sentence
* the label of the current word
* the label of the previous word
* and outputs a real-valued number

Next, assign each feature function a weight. Given a sentence s, we can now score a labeling l of s by adding up the weighted features over all words in the sentence:

QQ截图20171210211543

Finally, we can transform these scores into probabilities between 0 and 1 by exponentiating and normalizing.

Additionally, we divide training data randomly into two parts. The 80 percent part is used as the input of the CRF model, and the left part of training data is used to test the model’s output.

1. **Feature Engineering**

The features for the CRF model are listed in Table 3-2, which can be divided into two parts.

Table 3-2 Features

|  |
| --- |
| Basic Features  01: word wi  02: wi-1: the first left word of wi  03: wi+1: the first right word of wi  04: wi-2: the second left word of wi  05: wi+2: the second right word of wi  06: pi: the pos of word wi  07: pi-1: the first left pos of wi  08: pi+1: the first left pos of wi  09: pi-2: the second left pos of wi  10: pi+2: the second left pos of wi |
| Syntactic Features  11: dependencyi: the dependency parsing of wi  12: dependency\_parenti: the dependency parsing of wi’sparent  13: pos\_parenti: the pos of wi’sparent |

1. **Model Training**
2. We split the training data by 8:2 into training\_data01 and test\_data01.
3. Tag the instance sentence of training\_data01 according the themes and sentiment words: if the word is a theme, we tag it as T; if the word is a sentiment word, we tag it as S; if the word is a comma\, we tag it as “,” and others are O.
4. According the features, we constructed the template as follows

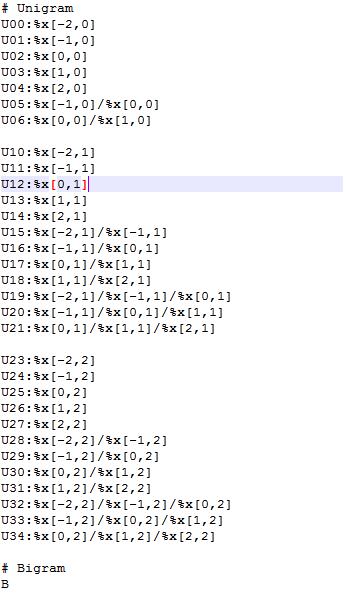


Figure 3‑2 The template of CRF

1. Model training: we use

*crf\_learn template traindata\_model.data model;*

to train the crf model and

*crf\_test -m model testdata\_model.data >output\_test.txt*

to test.

The iterations and the test instance are 277 as follows,

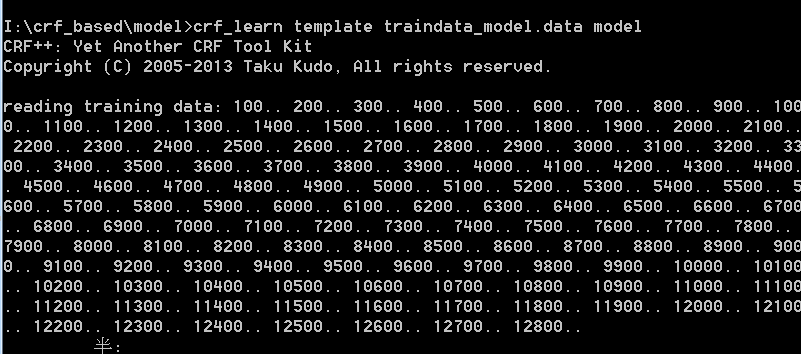


Figure 3‑3-1 The training model of CRF

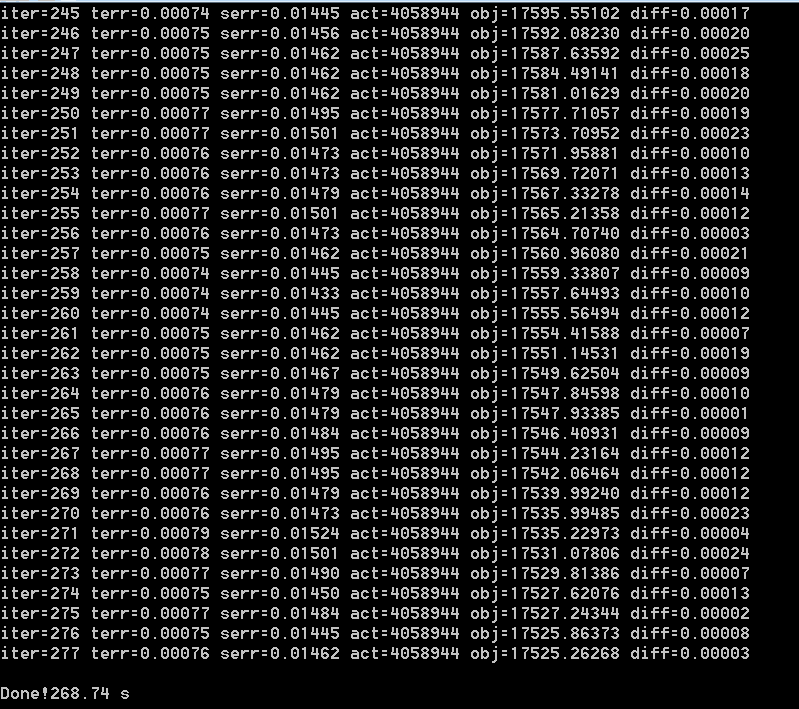


Figure 3‑3-2 The training model of CRF



Figure 3‑4 The test instance of tagging

**3.4.2 Rule Based**

Before identifying themes and sentiment words, we need to get the themes dictionary and sentiment words dictionary. We get the themes dictionary by reading the column of theme. Similarly, we get the sentiment words dictionary by reading the column of sentiment word.

Then, we identify themes and sentiment words by labeling each word in sentences. If the word can be matched with the theme dictionary, it will be labeled “T”. If the word can be matched with the sentiment word dictionary, it will be labeled “S”. The others will be labeled “O”. Therefore, the sentence will change into a labeled sequence.

## 3.5 The Matching of Themes and Sentiment Words and Polarity tagging

After segment and part-of-speech tagging, we get a sequence of tag. Next we need to pair theme words and sentiment words correctly.

The rule says, if sentiment word have no corresponding theme word or it’s not the first sentiment word about a theme word then set the corresponding theme word to ‘null’. That means every sentiment word have a theme word to pair with, but the theme word could be ‘null’.

After figuring out the rule, we observed the tagged data for a while and we found some pattern about pairing:

1. Theme word and sentiment word almost shows up in sequence. That means if a sentiment word shows after a theme word directly, we can pair them up easily.

*黑人/O 碳头/O 牙刷/T 好用/S -> (牙刷, 好用)*

1. If a sentiment word exists in a sentence first, it normally has no corresponding theme word which means we need to set the theme word to ‘null’;

*方便/S , 可挂/O 脖上/O , 当/O 项链/O -> (null, 方便)*

1. If a theme word follows another theme word directly which means no sentiment word exist between those theme words, the first theme word usually has no corresponding sentiment word. We can just discard it under the circumstances;

*纱布/T 材质/T 不错/S -> (材质, 不错)*

1. If a sentiment word follows another sentiment word directly which means no theme word exist between those theme words, the second sentiment word almost have the same corresponding theme word with the first sentiment word. According to the rule, we can just set the second sentiment word’s corresponding theme word to ‘null’.

*味道/T 清新/S 淡雅/S -> (味道, 清新), (null, 淡雅)*

We based on those patterns above to pair theme words and sentiment words. We get the result and observed on it for a while, we found that the theme words exist behind sentiment words many times which can’t be ignored.

*不方便/S 冲洗/T -> (冲洗, 不方便)*

There is a little bit of conflict between this circumstance and other patterns (mentioned earlier). To apply this circumstance to the data and keep earlier patterns unchanged at the same time, we add a new pattern as follow.

1. If there is exactly one theme word and one sentiment word in a sentence and the theme word exists behind the sentiment word, we just pair them up.

*物美价廉/S 东西/T -> (东西, 物美价廉)*

# Results



Figure 4-1 The result of preliminary



Figure 4-2 The result of final contest A



Figure 4-3 The result of final contest B

# 5. Conclusion

In this experiment, we did a lot work such as data processing, model training and so on. And we also suffered the encode issue which almost let us give up, but we figured it out finally.

In the process of implement, the hardest step is use CRF to tag out the theme words and sentiment words. And we tried many different features and different ways to train model. Since this is the first time we use CRF which is a powerful tool, we haven’t get the optimal result. In the future, we will learn more about it and hope we can use it proficiently to solve other problems next time.

To be honest, we learned a lot from the project. Firstly, we acquire some knowledge of NLP by finish this task. Then, we improve the programming level by several days’ coding. Thirdly, we can find solutions to solve bugs more efficiently.

# 6. Task Assignment

|  |  |
| --- | --- |
| Name | Task Assignment (code and report) |
| Leng Jia |  |
| Yan Hongyu |  |
| Gong XiaoChang |  |