A Method for Chinese Short Text Classification Considering Effective Feature Expansion

Mingxuan liu 1
School of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing 40065, China

Xinghua Fan 2
School of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing 40065, China

***Abstract*—This paper presents a Chinese short text classification method which considering extended semantic constraints and statistical constraints. This method uses "HowNet" tools to build the attribute set of concept. when coming to the part of feature expansion, we judge the collocation between the attribute words of original text and the characteristics before and after expansion as the semantic constraints, and calculate the ratio between the mutual information of the original contents and the features before expansion versus the mutual information of the original contents and the features after expansion as statistical constraints, so as to judge whether feature expansion is effective with this two constraints , then rationally use various semantic relation word-pairs in short text classification. Experiments show that this method can use semantic relations in Chinese short text classification effectively, and improve the classification performance.**

Keywords-component; short text; classification; semantic relations; semantic constraints; statistical constraints; HowNet.

I. INTRODUCTION

The short-text classification is an automatic classification for short texts (The text length is usually less than 160 characters). It is required for filtering information such as mobile phone short message, web comments, network chat, etc. And it has very important application prospect. However, short-text has many inherent characteristics such as short length, weak signal of concept description, etc. Therefore, traditional text classification methods cannot be applied to short-text classification better. An effective way of short-text classification is to use some extra information to assist classification [1-2]. Some methods has already achieved certain effect, one of which is based on hyponymy relation and is proposed by Sheng Wang[3] and another one is Agent and Patient Relation Acquisition for Short-text Classification which is proposed by Dingbang Wei[4]. Theoretically, when combining the word-pairs sets extracted by the above different two semantic relationships into a bigger word-pairs extension set simply, the extension result of short text will be better. However, experimental results in this paper show: when doing feature extension for test text by using the word-pairs extension set directly[5-6], the classification performance of short text isn't improved, but is slightly reduced. Analyzing the results, the reason is that when introducing word-pairs of a variety of different semantic relations at the same time, it will be at the greatly increased risk of the introduction of noise. The classification performance is reduced, for the features expansion isn't effective. If we want to expand features for short-text effectively, the following two issues must be resolved:1、How to determine whether noise is introduced when expanded, 2、How to apply different semantic relation word-pairs to classification of short text to improve the classification performance. For the above problems, this paper presents a classification method for short Chinese text considering effective feature expansion SCTCEFE(Short Chinese Text Classification Considering Effective Feature Expansion).

II. A CLASSIFICATION METHOD OF SHORT CHINESE TEXT CONSIDERING EFFECTIVE FEATURE EXTENSION (SCTCEFE)

The core idea of SCTCEFE is: through the constraint conditions, judge whether the knowledge expressed by the text before and after feature expansion changes, to ensure the availability of feature expansion and reduce the risk of noise introduction.

For the problem determining whether noise is introduced when expanded, by judging whether the knowledge expressed by extended text information and the original text are unified, solve the problem. Semantic relationship reveals the hidden semantic knowledge in the knowledge expressed by texts. Thus the article uses the following two methods to deal with different semantic knowledge: 1. relationship of similar concepts (Hyponymy relationship), by replacement of the similar concepts strengthen the description of the short text.

When using the concept information, we should determine whether the knowledge focus expressed by the original text is the similar concept between the relation word-pairs. for example, in the short text "宝马的外观非常重要", "外观" is the attribute of "汽车", we can determine the expansion is effective; But in the short text "我比较喜欢你那件宝马衣服的衣领", as "衣领" isn't the attribute of "汽车" , we can determine the expansion is invalid; 2. relationship of different concepts(agent-patient relationship), by introduction of new concept information mine the additional information of the short text[7].When using the information, we should judge whether there's strong correlation strength between new concept information and the other information of the original text.

For the problem of applying different semantic relation word-pairs to classification of short text to improve the classification performance, According to the different types of different relation word-pairs, the method establishes the semantic constraint for the similar concept relationship and the statistical constraint for the different concept relationship to judge whether feature expansion is effective.

The specific thread of SCTCEFE is: Build the attribute sets as the semantic constraint for the word-pairs of similar concept relationship, and calculate the ratio between the mutual information of the original content and the original feature versus the mutual information of the original content and the extended feature as the statistical constraint for the different concept relationship word-pairs. Combine with semantic constraints and statistical constraints; expand features effectively for short text with the use of relation words.

A. The standard of semantic constraint and statistical constraint

In this paper, we propose a concept of information integrity to describe the contribution value of a single word to the information integrity of a document. For example, the information integrity of a word A in the document D is described as P (A|D) :

$$P(A \mid D) = \frac{\sum_{0<i<m} I(A,W_i)}{I(m)} \quad (1)$$

I(m) represents the amount of information contained in the document D after the removal of a word A, m is the number of words in the document D, and I(A, Wi) represents the relevant information between a word A and the information Wi of a certain word in the text D. In this paper, we suppose that every sentence is an independent knowledge-expressed document. By feature expansion, we use $consist\left[\left(A,\bar{A}\right) \mid D\right]$ to express the consistency of text-expressed knowledge after expansion. When $consist\left[\left(A,\bar{A}\right) \mid D\right] > 0$, there is a certain consistency between the knowledge of the extended text and the original text.

Semantic constraint: used for judging whether the similar concept information of the similar relationship word-pair is the keystone in the text. When there appears an attribute word which can match the relation word-pair simultaneously in one text, we can suppose that the similar or identical concept information of the word-pairs is used in the text, which can be expressed as:

$$consist\left[\left(A,\bar{A}\right) \mid D\right] = \frac{P\left(\bar{A} \mid D\right)}{P\left(A \mid D\right)} = \frac{\sum_{0<i<m} I(\bar{A},W_i)}{\sum_{0<i<\sum_{0<i<m} I(A,W_i)} I(A,W_i)} \quad (2)$$

$\bar{A}$ represents the relation word of concept information similar to the word A. When there is an attribute-host relationship between an attributive (or attributive value) word and the word $\bar{A}$, then $I(\bar{A},W_i)=1$, otherwise it is 0. In this paper, when $\sum_{0<i<m} I(A,W_i)$ =0, $consist\left[\left(A,\bar{A}\right) \mid D\right]$=0.

statistical constraint: used for judging whether the word-pair which has different concept information is the hidden information of the text, when the ratio between the mutual information of the introduced new concept information and the original text versus the mutual information of the original concept information and the original text is within a certain range, it's considered that the concept information hides in the original text, which can be expressed as:

$$consist\left[\left(A,\bar{\bar{A}}\right) \mid D\right] = \frac{P\left(\bar{\bar{A}} \mid D\right)}{P\left(A \mid D\right)} = \Phi - \left| \frac{\sum_{0<i<m} MI\left(\bar{\bar{A}},W_i\right)}{\sum_{0<i<m} MI\left(A,W_i\right)} - 1 \right| \quad (3)$$

$\bar{\bar{A}}$ is a word set which includes the new concept information and is associated with the word A. $MI\left(\bar{\bar{A}},W_i\right)$ is the mutual information between the word $\bar{\bar{A}}$ and the word W_i, and Φ is an adjustable threshold.

1) Establishing the attribute sets of concept information

Understanding of attribute: Any object must carry a set of attributes. Similarities and differences between the objects are determined by the attributes they each carry. There will be no object without attributes. Human beings have natural attributes such as race, color, gender as well as social attributes such as nationality, class origin, job, wealth etc. Under specific conditions, it is true to say that the attached attributes are even more important than the host itself. This paper presents a method by extracting the co-occur word-pairs which can constitute the relationship of concept-attributes to establish the attribute sets of all concept information in the training set.

A conventional method of Getting concept - attribute relations: HowNet determines all the hosts corresponding to attributes, and is signed with pointer &, for example: 外观: attribute|属性,appearance|外观,&physical|物质;This concept indicates that appearance belongs to the information of "属性", it can be understood as "外观", relying on its host "物质". On the other hand, the hyponymy relation of words indicates the universality and individuality between words. On the basis of universality inherits from higher word, the lower word has part of its own characteristics. Therefore, to any word-pair (C1,C2), if the first semanteme of C1 is an attribute, and its host is C2 or the higher concept of C2 , then C1 and C2 constitute the concept-attribute relationship directly, and C1 exists as the attribute of C2.

An improved method: because Chinese short texts often omit attribute words and adopt 'concept + attribute value' expression in order to give the essentials in simple language. Through common methods we cannot affirm the concept-attribute relationship between the two. However, in practice, attribute value embodies attribute, which should also be summed up in the attribute set of concept. This paper introduces an improved method to obtain concept-attribute relationship word-pairs:

a) For any word-pair（C1,C2）, if it contains one attribute value word, execute step (2), if it contains one

attribute word, execute step (3)，if it contains neither, the word-pair has no concept-attribute relationship；

b) Find attribute words corresponding to the attribute value words, then execute step(3);

Find the host of the attribute word, and judge whether the other word in the word-pair is the lower concept of the host, if yes, the word-pair has concept-attribute relationship, or else the word-pair has no concept-attribute relationship.

2) Building characteristics word-pairs and the attribute sets of the word-pairs

For feature word-pairs which have a similar concept relationship, in order to calculate its semantic constraint, we need to construct the attribute sets of the relationship, Take the hyponymy relations adopted by this paper for example. The feature of the attribute set applied to these feature word-pairs is constructing the concept-attribute relations with both the words in the same pair.

c) according to the method of literature [3], obtain feature word-pairs set I of hyponymy relation, I1={(W1,W2) | (W1,W2) appearing in the word-pairs obtained from the training set, and constitute the hyponymy relationship };

d) To every word-pairs (W1, W2), according to the method mentioned in 1.1 in this chapter, obtain the attribute set of W1, I11= {(C1) | (W1,C1) appearing in the word-pairs obtained from the training set, and (W1,C1) constitute the concept-attribute relation}. Also the attribute set of W2, I12 = {(C2) | (W2,C2) appears in the word-pairs obtained from the training set, and (W2,C2) constitute the concept-attribute relations};

e) Construct attribute sets of every feature word-pairs (W1, W2), I2 = {(C) | C∈I11 且 C∈I12}.

B. The Algorithm Description of Feature Expansion based on semantic relations

Input: short-text for test, feature word-pairs sets, attributes sets of feature word-pairs, mutual information between the words.

Output: Feature vector of the extended short-text.

1) For any word of the short-text, inquire the word-pairs sets. if there exists a record of a similar concept relationship, go to step 5, else if there exists a record of different concept relationships, go to step 2, else go to step 9;

2) If there is only one word-pair ti-tj, go to step 4, else if there are several word-pairs, go to step 3;

3) Extract the right words of all the word-pairs related to ti and form into Tx, if ∋tj∈TX, and tj can be found in the vector space of this short-text, go to step 7, else extract tj , the right word of the word-pair with the highest strength, and go to step 4;

4) Extract tj ,the right word of this word-pair, if tj cannot be found in the vector space of this short-text, go to step 6, else go to step 7;

5) Extract the word set TY in the text, if there exist tk∈TY and tk∈TZ（attribute set of the word-pair（ti，tj））, go to step 8, else go to step 10;

6) Calculate the mutual information between tj and other words in the text, and go to step 8 when meeting the requirements , else go to step 10;

7) Calculate the mutual information between tj and other words in the text, and go to step 9 when meeting the requirements , else go to step 10;

8) Insert tj into the vector space of this short-text;

9) Raise the frequency of tj in the vector space of this short-text at λ. (0<λ<1);

10) Don't extend this word, and input and seek the next word.

III. EXPERIMENT

When using the STCEFE with the use of statistical constraints and semanteme constraints, the experiment to test the effect of the short-text classification uses two different semantic relations.

Experiments dataset: The dataset used here is 4702052 Chinese short-texts which compose of 12 categories, including finance, real estate, international news, national news, military, science and technology, women, cars, book reviews, sports, games, entertainment. All of short-texts are titles of the news or Netizen comments about topics from Sina website and Netease website. We divide texts of each category into four parts randomly and averagely, one part as testing data, the rest as training data.

The Evaluation of Classification Performance: In this experiment, we use the following indexes in the evaluation of classification: Precision (P), Recall (R), F1-measure, and Macro-F1= $\frac{1}{n}\sum_{i=1}^{n} F1_i$

In experiments, we only adopt statistical constraint when selecting 4 kinds of features respectively. According to different thresholds, the revealed change trends of F1 are in nearly identical form, and when threshold Φ is set as 0.1, the classification result will work out for the best when Hyponymy Relationship word-pairs and Agent and Patient Relation word-pairs are both used for text classification. However, the result is still not ideal, we can conclude that only by calculating mutual information the response to whether the introduced new information matches the other information in the original text is not so accurate, thus the method should be in conjunction with other methods.

Experiment introduction: By taking sentences in the text as windows, extract the binary word-pairs in the training set and calculate all mutual information between every word-pair. Build the attribute set of concept according to the method in 1.1 of this article, using the hyponymy word-pairs and agent-patient word-pairs to build an attribute set of feature word-pairs by the method in 1.2 of this paper.

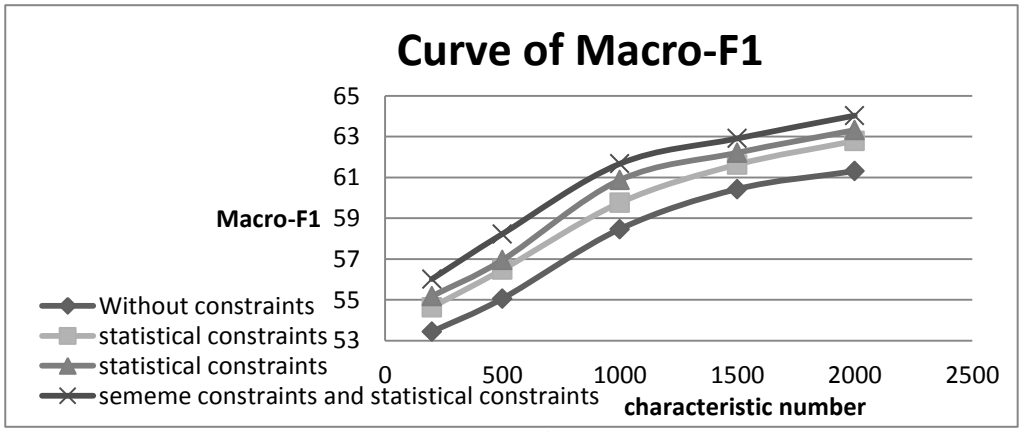

Figure 1. Curve of Macro-F1 by experiment

This experiment uses CHI feature selection, and takes the Naïve Bayes as the classifier. Then 5 groups of experiments for short-text classification are taken, and the sizes of the feature space are 200, 500, 1000, 1500, 2000.

Method 1: According to the expansion method described in [5], expand test text set by hyponymy word-pairs and agent-patient word-pairs.

Method 2: According to the expansion method only by the statistical constraint in this article and threshold is set as 0.1, expand test text set by hyponymy word-pairs and agent-patient word-pairs.

Method 3: According to the expansion method only by the semantic constraint in this article, expand test text set by hyponymy word-pairs and agent-patient word-pairs.

Method 4: According to the expansion method both of the semantic constraint and the statistical constraint in this article and threshold is set as 0.1; expand test text set by hyponymy word-pairs and agent-patient word-pairs.

The experimental results:

The classification results are shown in the figure, and we can arrive at the following review:

A. *Method 2 and 3 are both superior to method 1, indicating that when many kinds of semantic relationship words are used at one time to assist in categorizing the text, the traditional classification performance is far from ideal. By the expansion strategy judging the consistency between the expanded contents and the original text to decide to do feature expansion or not, method 2 and 3 could bring a certain improvement on the categorizing performance.*

B. *The result of method 3 is slightly superior to method 2, indicating the help of semantic constraint to the accuracy of text expansion is slightly superior to the help of statistical constraint.*

The result of method 4 is the best, indicating when using many semantic relationship words to assist in categorizing the text, a combination of semantic constraint and statistical constraint could be a bigger help to text categorizing. But the categorizing performance is still not high, indicating there is still a big part of hidden information to be mined in the short text.

IV. Conclusion

The paper proposes a short text classification method in consideration of characteristic effective extension, and constructs a concept of information completeness through semantic constraint and statistical constraint. Based on the concept when doing feature extension for the original text with the use of semantic relation word-pairs, judge whether the expansion is valid. Moreover, review the effectiveness of the classification method through experiments, and thus conclude that: 1) the short text classification method combining many semantic relationships and doing effective expansion with considered characteristics could enhance classification performance; 2) We can consider the combination of the other methods, to acquire more accurate constraint information judging the consistency of text contents.

Acknowledgment

The research is supported in part by the National Natural Science Foundation of China under grant number 6070301, cultivate project for significant research plan of the National Natural Science Foundation of China under the grant number 90924029 named “Research on the mechanism and related technology of public opinion on internet for abnormal emergency” and the Nature Science Foundation of Chongqing province in China under grant number CSTC, 2009BB2079.

References

[1] ZELIKOCITZ S.Transductive LSI for short text classification problems[C]//Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference. Florida: AAAIPress, 2004:556-561.

[2] ZELIKOCITZ S, MARQUEZ F. Transductive learning for short text classificaton problems using latent semantic indexing[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2005, 19(2):143-151.

[3] Sheng Wang, Xinghua Fan , XianLin Chen. Chinese short text classification based on hyponymy relation[J]. Journal of Computer Applications, 2010, 30(3):603-606.

[4] Xinghua Fan, Dingbang Wei. A Method of Agent and Patient Relation Acquisition[C].2011 International Conference on Computer Science and Informatin Engineering.

[5] Xiwei Wang, Xinghua Fan, Jun Zhao. Method for Chinese short text classification based on feature extension[J]. Journal of Computer Applications, 2009, 29(3):0843-0845.

[6] Xinghua Fan, Maosong Sun. A High Performance Two-Class Chinese Text Categorization Method [J].Chinese Journal of Computer.2006, 29(1):124-131.

[7] Ramage D,Rafferty AN,Manning CD.Random walks for text semantic similarity//Proceedings of the 2009 Workshopon Graph— based Methods for Natural Language Processing.Suntec,Singapore,2009:23-31.

AUTHORS PROFILE

Mingxuan Liu born in 1988, is Graduate of Chongqing University of Posts and Telecommunications. His research interests is Chinese information processing.

Xinghua Fan born in 1972, is Professor and Ph.D. of Chongqing University of Posts and Telecommunications. His research interests include natural language processing and information retrieval.