A Construction of Emotion Thesaurus Basing on Chinese Character and Empirical Knowledge

Yu Zhang[1], Zhuoming Li[1], Fuji Ren[1] and Shingo Kuroiwa[1]

Faculty and school of Engineering, the University of Tokushima, 2-1 Minamijosanjima-cho,Tokushima 770-8506, Japan
{zhangyu, ren, kuroiwa}@is.tokushima-u.ac.jp

Abstract. This paper presents an effective and practical approach on the emotional lexicographic component of Chinese documentation and metadata specification. In this paper, based on Chinese character and empirical knowledge, we construct an emotion thesaurus in which there are some elements recorded, such as image value, word frequency and affective label which including emotion intensity and affective label and so on. Additionally, a sentence corpus are developed to improve the performance of the emotion thesaurus. Finally, we discuss the advantage of the emotion thesaurus and according to the evaluation experiment this thesaurus produces much better results.

Keywords: Emotion thesaurus (ET), affective label, sentence gatherer, emotion intensity

1 Introduction

Textural information is a huge source of emotion such as words and emotions. Our group is working on emotion extracting from textural information. Emotion can be clearly realized by people from a novel or a story, Researchers first will do lexical analysis then do syntax analysis, and from the keyword we got, we give the emotional information to the keyword using emotion thesaurus we have made.

Why to develop the emotional thesaurus? It is proverbial that Chinese is difficult for non-native speakers, such as using words to make sentences even though they know each word. Because the difficulty is from the descriptive word. When a descriptive word appears in different context it may express different meanings. Every descriptive word have it's own tendency, negative or positive. When using these words people have to know the tendency before using them. Some words are used as politeness and honorifics, while some words are impolite or discriminate and should not be used in friendly commercial correspondence. Along with the international exchanges getting more and more tightly, the request to culture and literalness of all countries also raises, it seems the traditional way is not suitable for the developing condition, by which words are translated only by likeness, and similarity in shape. Although we can't enumerate in detail the civilization for several thousand years of every word, we can consider add the emotional attribute to the vocabularies. That is

Advances in Artificial Intelligence and Applications
Research in Computer Science 32, 2007, pp. 330–340

Received 12/06/07
Accepted 31/08/07
Final version 21/10/07

considered to be very necessary. Some emotion dictionary have already made in Japanese. [1][2]

For example, "he got praise from his teacher" if do not know the different of "受到/shou dao", "遭到/zao dao" foreigner always make mistakes using "他遭到了老师的表扬" instead of "他受到了老师的表扬". The point on this argument is ignoring the emotional attribute of one word, when the communication between countries become more tightly the understanding of culture and linguistic is becoming important. While in Chinese, regretfully there is no one dictionary containing clear emotional information. Emotion studying and analyzing in Chinese text is still on a starting stage. Lacking of a comprehensive electronic Chinese thesaurus is the crux. There are insistent demands of an emotional thesaurus for more help. Hence, in our study we build up our own corpus and construct an emotional thesaurus. And in this paper we will introduce our work about the construction of an emotion thesaurus (ET).

This paper is organized as follows:

We present the related work and the necessity of emotional thesaurus in our study in section 2. In section 3, we talk about the construction of the emotion thesaurus (ET) detailedly, such as the vocabulary entry, image value, affective label. In the part of affective label we introduce the method of distributing emotion category using Chinese character and empirical knowledge. The section 4 is the evaluation based on the result of investigating experiments. In last section 5 we remark the conclusion and future work.

2 Related Work

Much research has been done on how to classify emotions. In the field of psychology and linguistics, some researchers have done some work with emotion such as emotion classification from facial expression [3], emotion analysis. [4]. In these research emotion are mostly divided into 4-8 kinds. Such as Ekman who had classified emotions into 6 kinds depending on the facial expression, they are happiness, sadness, fear, disgust, anger, surprise considered to be the basic emotion of human beings. To this basic emotion set love was added, since it has been extensively promoted as a basic emotion with the realm of prototype approaches [5]. In some Japanese dictionary [2], there are eight kinds of emotion recorded with enough examples.

In some archaic Chinese literary work, 7 kinds of emotion are often used from a mass of textual recordation. They are "disgust, happiness, anger, sadness, fear, love, and desire". In psychology and common use, emotion is an aspect of a person's mental state of being, normally based in or tied to the person's internal (physical) and external (social) sensory feeling. Some Chinese psychologist have done deep research about emotion, mood of human, they get out their classification basing on point of view of psychology. In contemporary Chinese the emotion word based on psychology and susceptibility can be divided into 24 kinds. [6]

But in our research, the more kinds of emotion don't mean the better the result is, and we must choose the primary emotion as our subject. Comparing the basic emotion

of Ekman's [7], at last we add 6 kinds of emotion depending on the word frequency from 24 kinds and plus (the) equable.

Table 1. 13 kinds of emotion.

Emotion						
Happy Love	Sad Expectant	Fearful Nervous	Disgusted Regretful	Angry Praiseful	Surprised Shy	Equable

3 ET

We have constructed a corpus of emotional words called emotion thesaurus which including about 3700 words selected from two dictionaries and one corpora.p[8][9]

Based on the statistics result in People's Daily tagging corpus of the Institute of Computational Linguistics of Peking University, the word's emotional trends are described and formalized in our dictionary. In order to construct the database of emotional information we select phonetic, part of speech, image value, and the affective label including category of emotion and emotion intensity to compose the items of the thesaurus as shown in Figure 2.

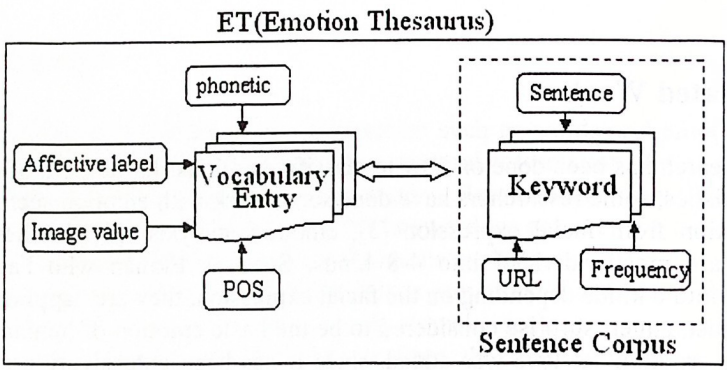

Figure 2. Structure of ET

3.1 Vocabulary Entry

Vocabulary entry is an entry for a word with an emotion-related meaning or connotation. e.g. ,沮丧(gloomy),欣喜(delighted),哭泣(cry)，愿望(desirability), etc. In our ET, we reserve almost all the entries of adjectives, verbs and nouns which denote affects directly. Others like adverbs, some of them with no affect (such as Amplifier words only represent degree) when they are used with some affective words but can strengthen the intensity of words, However, itself has no affective meaning. Words like these, we only give intensity to them. Obviously, POS (part of

speech) is interrelated with every entry and is momentous in emotion judging. There are about 3700 vocabulary entries were gathered in this ET.

3.2 POS

POS is the term used to describe how a particular word is used. All words in the ET are made of nouns, adjectives, verbs, adverbs and also some phrases. From a mass of investigations from Chinese dictionaries [8], the distribution of POS in all of Chinese words is like following.

In modern Chinese, "normal" nouns (i.e. not "time words" or "nouns of locality") make up 53.0% of all vocabulary

Verbs make up 20.1% of all vocabularies

Adjectives make up 4.5% of all vocabularies

Adverbs make up 2.0% of all vocabularies

Phrases make up 8.6% of all vocabularies.

Noun word is used to identify a person, thing, animal, place, and abstract idea of affectively. Adjective word is one of the most important parts of the Chinese Grammar. It is used in describing, identifying a Noun or a Pronoun with affect. A Verb expresses certain actions, events, or states of being. Some emotions are expressed by these words. Not only affective but also some nonaffective words are selected into this ET because the intensity of degree is needed.

那些孩子们欢欣雀跃着，看起来非常开心。----------"非常"（very）

(Those children are gamboling and seem very happy.)

Some words with more than one POS are also gathered as individual entry in the ET. For example:

小心前面的车---------- " 小心 " (beware of) verb. Nervous

他做事很小心---------- " 小心 " (careful) adj. Praiseful

3.3 Image Value

Sometimes a non-native person has trouble in understanding Chinese or misunderstand the meaning of Chinese in listening and speaking. The sticking point is that some words in Chinese have its tendency of + or -, which makes the word own its attitude, such as appraisement, positive and negative. Additionally, for some application of the ET, we find the appraisement of the word is needed to add as the complementary property.

In our ET, there is another intensity of emotion called image value used to scale the plus or minus image of the emotion word.Many authors agree that emotions can be organized roughly into a two-dimensional space whose axes are evaluation (i.e. how positive or negative the emotion is) and activation (i.e. the level of energy a person experiencing the emotion is likely to display)[4]

According to the dictionary of "chang yong baobianyciyu xiangjie cidian", 1103 words are recorded in it, and each word is tagged by the appraisement. This makes our ET more effective. For example, "surprise", in some situation such as, "get a good news", the "surprise" will be positive, while in other situation such as, "I can not believe I fall the exam, it really surprised me" that points to a negative meaning. Worse than negative or better than positive we defined it as derogatory or commendatory and between them we use neutral as a median. Here are the five levels we have defined:

derogatory-> negative-> neutral-> positive-> commendatory

Each word was rated for tendency (positive or negative) and for the degree which presents a given emotion on a five-point scale ranging from –2 to +2 by increments of 1.

3.4 Sentence Corpus

Language corpora usually represent a large collection of representative samples obtained from texts covering different varieties of language used in various domains of linguistic activities. Corpora can be looked as the abbreviation of CORPORA - Capable of Representing Potentially Unlimited Selection of Texts.[10]

The implementation of complex multimedia lexica in hypertext formats is potentially a task of extremely high complexity. It is suggested that as a preliminary step towards designing electronic lexica of various kinds, including Web lexica, a requirements specification in terms of first principles of the lexicon sciences is needed. Electronically-made corpora are new things, and we are yet to reach a consensus to what counts as corpus and how it should be classified.

In this study, we classify the corpora in a very general way, which way focusing on the types of need in the present of Chinese context. They contain written or spoken text, new or old, long or short, which can be obtained from Internet. Obviously the contents are extensive, which can be stories, news, speech, and even they can be extracts of messages of varying length shown in figure 3.

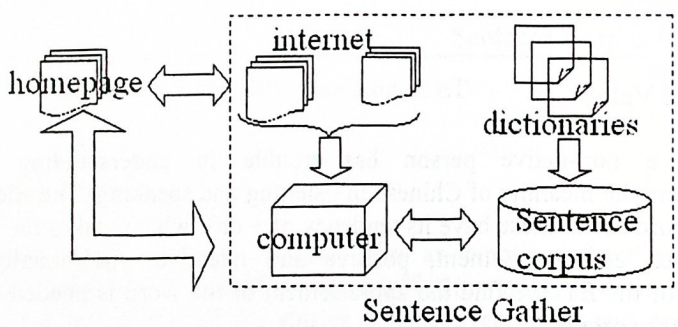

Figure 3. The flow of construct sentence corpus.

In emotion tagging, we consider that one affective word has more than one emotional category, hence the conformation specially includes 13 kinds of emotion which we have defined besides other components. We gather 50 sentences for per

keyword elementarily as the testing text, then distribute the emotion category to the word in one of these 50 sentences.

It is a very tremendous work to give each word tagging of 13 kinds emotion. The development of Internet brings more and more facilities to the people. Everyday there are a fold of textual information gushing from everywhere just like an enormous storehouse of natural language, and it also provides us a lot of usable resources we can depend on. The other advantage of this way is ensuring the functionality of the source. Hence, we exploit a sentence gathering tool to replace gathering sentence from Internet randomly.

3.5 Affective Label

Affective label is a tag of emotion attribute and the emotion intensity. Sometimes the word is not in only one of the 12 categories of emotion, it is in two or three kinds of those emotions. In such an instance every emotion is recorded in our dictionary.

Distributing emotion category. In our study, there are much more detail progresses in the ET. In this ET there are 13 kinds of emotion provided to each word and also the numerical intensity indicated. Such as,

"忧郁"（somber）---adj.---sad 0.36

"忧伤"（somber）---adj.---fearful 0.22

"忧伤"（somber）---adj.---nervous 0.42

Any affect given word may have multiple entries in the ET, differentiated by part of speech and emotion category.

In order to build the ET, there are some preliminary works needed to do such as to indicate the numerical intensity. It is obvious that the emotion has itself ambiguous meanings and sometimes the emotion may puzzle the people. Hence, we use empirical knowledge and method to identify the strength of each emotion to repair the emotion category of words.

Method. Despite its phenomenal salience, (where is the definition of intensity) the intensity of emotion remains a neglected issue of emotion research, and in this study we focus on the intensity of emotions. Emotion intensity represent the strength of affect degree ranging from 0 to 1. In previous publication, there is not a specific criteria of the emotion study, contrarily emotion study is empirical and collect data-sampling depending on the experiment rather than on the elaborate rules.

For a long period in the linguistic study, the empirical data is utilized in lexicography. Samuel Johnson used examples from literature to illustrate his dictionary (1755), while the Oxford English dictionary used citation slips to study and illustrate the usage of words in English. Today, corpora are used for dictionary preparation as they play important role in dictionary building. [11] Moreover, the empirical approach to language study has been identified to be more reliable and authentic than rationalistic (based on intuition) approaches. [12]

In fact, we get the numerical scale of each emotion ascribing to each affective entry from the metadata---sentence. It is a contribution of an empirical database in the form

of corpora. The criteria for scaling emotion intensity is the combination of empirical experiential marking and statistical method.

Here the emotion category of a word can be one or more than one. There are 13 kinds of emotion (shown in Figure 6 and Table 3), which of the word are restricted in these categories. We manually mark the affective word with the emotion category. If there is somewhat any emotional information of the 12 emotions (ex equable), we give the value "1", otherwise the value is "0". Sometimes we can not guarantee that the category is able to be given strictly to the word during the estimation, because the diversified ambiguities inhibit the real emotion. In the Table 3, E13 (equable) is utilized to detract the intensity from other emotions (E1, E2, ..., E12). It means that sometimes the word can be considered as an equable word without any emotion information. In our study, this action is used to get the exact proportion of the emotion.

Table 3. An example of how to mark the emotion. Remark: E1=happy, E2=sad, E3=fearful, E4=disgusted, E5=angry, E6=surprise, E7=expectant, E8=love, E9=nervous, E10=regretful, E11=praiseful, E12=shy, E13=equable

Keyword	No.	Sentence	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13
悲痛	1	东盟 10 国和中国、日本、韩国的外长对以军袭击联合国观察哨所事件感到震惊和悲痛，并向受害者家属表示慰问。	0	1	0	0	1	1	0	0	1	0	0	0	0
悲痛	2	对于刘老师的不幸逝世，我深感十分悲痛，我要化悲痛为力量，接过刘老师高举的伟大红旗。	0	1	0	0	0	1	0	1	0	0	0	0	0
悲痛	..	....	..	..	..	..	..	..	..	..	..	..	..	..	..
悲痛	50	.....	0	1	0	0	0	0	0	0	0	0	0	0	0

We experiment with these 50 sentences in this paper and get the possession of each emotion attribute to these 13 kinds of emotion. Using them as the basic contextual data, we extracted emotion attributes from them at last. (Figure 6).

	13 kinds of emotion	
word	E11 E12 .....E1j..... ... E113	S1
	E21 E22 .....E2j..... ... E213	S2
	....	...
	Ei1 Ei2 ..... .Eij..... .. Ei13	Si
	....	...
	E501 E502 ...E50j..... E5013	S50
Sum(E)	E1 E2 ... . Ej..... ... E13	

S=Sentence

Figure 6. The matrix of the emotion calculation, including 50 sentences and 13 kinds of emotions. Eij (i = 1, ..,50, j = 1, ...,13) is the value "1" or "0" to denote that the word has the emotion information or not. Si (1 = 1, ..., 50) is the number of the sentence.

In these 50 sentences the total of each emotion is $E_j = \sum_{i=1}^{50} Eij$. For each kind of emotion we get a sum, and share it to the 50 sentences, then the average of each

emotion is $Ej_{ave} = \frac{E_j}{50}$. That is considered to be the affective intensity of the 13 emotions of the word. Here Ej_{ave} is a center centage of every emotion distribution in sentences for each word. (The result is shown in the Table 4.)

Table 4. The result of 50 sentences emotion calculation.

Word	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13
悲痛	0.00	0.61	0.01	0.00	0.04	0.11	0.08	0.06	0.01	0.01	0.00	0.00	0.07
开心	0.90	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.06
伤心	0.00	0.51	0.03	0.08	0.03	0.02	0.00	0.11	0.14	0.06	0.00	0.00	0.02
忧愁	0.00	0.37	0.28	0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.01
难过	0.00	0.47	0.03	0.09	0.08	0.05	0.02	0.00	0.17	0.01	0.00	0.03	0.04
忧郁	0.00	0.27	0.17	0.07	0.01	0.00	0.00	0.01	0.31	0.00	0.01	0.00	0.16
忠厚	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.50	0.00	0.44
勇敢	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.08	0.00	0.00	0.41	0.00	0.44
充沛	0.04	0.00	0.00	0.01	0.00	0.00	0.05	0.04	0.00	0.00	0.32	0.00	0.54
...	...	...	...	...	...	...	...	...	...	...	...	...	...
贪婪	0.03	0.03	0.01	0.39	0.14	0.05	0.00	0.08	0.01	0.00	0.00	0.00	0.26

Furthermore, this ET is not only a static dictionary but also a mutual one, we also design a new function to make it active. Provision for its constant growth will reflect the meaning changes occurring in the language. As a result, more and more new words are put into use while some uncommon words are out of use, and some of the words will be recurrently used with the new values. Gradually, over time they will achieve a diachronic dimension representing data obtained from a wider range of time. Thereby we make our sentence corpus grow progressively and the sentences are gathered and keep them updating momentarily. Such corpora will help us to identify new words and phrases, to locate newly coined technical terms, even to know actual date of coming of words to track variation in usage of lexical items and phrases, to observe changes in meanings of words, to follow changes in sentence structures, and to reconstruct the emotion categories of words. At last, the intensity of emotion will be inclined to exactness gradually. The sentence gatherer will automatically update every month. For the new words, we also make a participant IE page to collect informational context of new words, as shown as Figure 5.

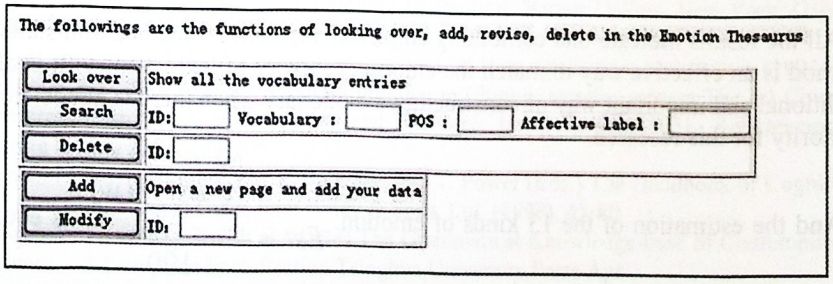

Figure 5. Interface of data updating of ET

4 Evaluation

In order to avoid the category estimation becoming arbitrary and straying somewhat from the strictly affect domain, we invite 14 undergraduate students of the University of Tokushima to participate our experiment, whose native language is Chinese. Each participant is given 100 words with the emotion distributed randomly which are selected from the emotion thesaurus in this paper. The participant is requested to estimate whether the given words with the defined 13 kinds of emotion are compatible to their image or not. If not, the participant will be asked to estimate the correctness of each entry comparing to the expected value. The mark will be a 10-point scale and the decimal digits were limited to 1.

In the evaluation experiment, each word is marked by an emotion value, therefore for each word we get 14 values from all of the participants. We use equation (1) to calculate average evaluation result of 14 participants. Here we remove the max and min marked values, and then the left 12 values are averaged. Figure 7 shows the average evaluation result gotten from 14 participators and 100 words.

$$Est_{ave} = \frac{\sum_{a=1}^{14} Est_a - Est_{\max} - Est_{\min}}{12} \quad (1)$$

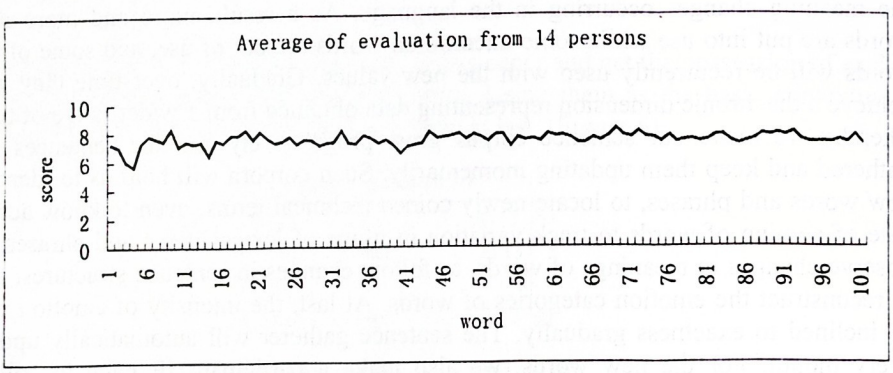

Figure 7. Average of evaluation from 14 persons

All the results indicate that combining empirical experiential method and statistical method is an effective way to match the emotion category of lexical entry. It is also a traditional and important way of constructing a dictionary when there is almost no any authority for this research.

And the estimation of the 13 kinds of emotion $Est = \frac{\sum_{w=1}^{100} Est_{ave}(w)}{100}$ is equal to 7.54 matching to the level D (6-8: somewhat agree). This evaluation result shows that

the emotion value defined in our ET is acceptable and the ET we build in this study is reliable.

5 Conclusion

In this paper, we discussed the current research situations on affective computing according to emotion dictionary and proposed an effective and practical emotion thesaurus based on empirical knowledge and the character of Chinese word to make the emotion definition much more accurate. During the construction of emotion thesaurus, (add something about the content of the thesaurus refer to) we also developed an automatic sentence gathering tool named sentence gatherer to assist to reduce the workload of improving the accuracy of emotion thesaurus. After the automatically sentence collection and manually labeling processing of the emotion words in the selected sentences, the emotion thesaurus construction reconstruction process is terminated. In order to evaluate the ET, we invited some students whose native language is Chinese to evaluate the ET in our evaluation experiment. From the final result, a good correspondence between affect sets and human judgments was discovered. It can be concluded that the emotion estimation in our study is reliable and represents the significant of our study. The ET is proved to play an important role in lexical software application with reusability, interoperability, and portability.

Because the sentence corpus is getting extended with our automatic gathering tool, the ET will be updated momentarily. In future, we will apply it into an emotion classification model as a part of corpus for further research.

Acknowledgments. This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300029, 17300065, Exploratory Research, 17656128.

References

1. Ronald Suleski and Masada Hiroko Affective expressions in Japanese : a handbook of value-laden words in everyday Japanese, Tokyo Hokuseido Press, (1982)
2. Yitirou Hiejima : a short dictionary of feelings and emotions in English and Japanese, (1995)
3. Ekman, P., Friesen, W.V., and Elsworth, P.: Emotions in the Human Face, London: Cambridge University Press, (1982)
4. Plutchik, R. : The psychology and biology of emotion. Harper Collins, New York, (1994), p.58
5. Fher, B. and Russell, J.: The concept of love viewed from a prototype perspective. (1991)
6. Xiaoying Xu, Jianhua Tao : Emotion dividing in Chinese emotion system The 1st Chinese Conference on Affective Computing and Intelligent Interaction(ACII'03), 8-9 December, Beijing China, 2003, 199-205
7. Ekman, P., Basic Emotions In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons, Ltd. (1999), 45-60
8. Shiwen Yu, Xuefeng Zhu et al, (1998) The Grammatical Knowledge-base of Contemporary Chinese – A Complete Specification Tsinghua University Press Apr

9. Xianzhen Guo, Chang Yong (1996) ChangyongBaoBianYiXiangJie dictionary The Commercial Press
10. Dash, Niladri Sekhar and B.B. Chaudhuri: Relevance of Corpus in Language Research and Application, International Journal of Dravidian Linguistics. Vol. 32. No. 2. (2003), 101-122.
11. Mindt, D An Empirical Grammar of the English Verb: Modal Verbs,1995
12. Eric Brill and, Raymond J. Mooney: An overview of empirical natural language processing AI Magazine Contents, Volume 18(4): (1997) Winter, 13-24
13. Bates, J. The Role of Emotion in Believable Agents. Communications of the ACM.37(7), (1994), 122-125
14. Mihalcea, R., and Liu, H. A Corpus-based Approach to. Finding Happiness. In Proc. of the AAAI-CAAW'06, (2006)
15. Zhan Weidong, Chang Baobao, Dui Huiming, Zhang Huarui, Recent Developments in Chinese Corpus Research, The 13th NIJL International Symposium, Language Corpora. Tokyo, Japan. (2006)