Anomaly Detecting within Dynamic Chinese Chat Text

Yunqing Xia
Department of S.E.E.M.
The Chinese University of Hong Kong
Shatin, Hong Kong
yqxia@se.cuhk.edu.hk

Kam-Fai Wong
Department of S.E.E.M.
The Chinese University of Hong Kong
Shatin, Hong Kong
kfwong@se.cuhk.edu.hk

Abstract

The problem in processing Chinese chat text originates from the anomalous characteristics and dynamic nature of such a text genre. That is, it uses ill-edited terms and anomalous writing styles in chat text, and the anomaly is created and discarded very quickly. To handle this problem, one solution is to re-train the recognizer periodically. This costs a lot of manpower in producing the timely chat text corpus. The new approaches are proposed in this paper to detect the anomaly within dynamic Chinese chat text by incorporating standard Chinese corpora and chat corpus. We first model standard language text using standard Chinese corpora and apply these models to detect anomalous chat text. To improve detection quality, we construct anomalous chat language model using one static chat text corpus and incorporate this model into the standard language models. Our approaches calculate confidence and entropy for the input text and apply threshold values to help make the decisions. The experiments prove that performance equivalent to the best ones produced by the approaches in existence can be achieved stably with our approaches.

1 Introduction

Network Informal Language (NIL) refers to the special human language widely used in the community of network communication via platforms such as chat rooms/tools, mobile phone short message services (SMS), bulletin board systems (BBS), emails, blogs, etc. NIL is ubiquitous due in special to the rapid proliferation of Internet applications. As one important type of NIL text, chat text appears frequently within increasing volume of chat logs of online education (Heard-White, 2004) and customer relationship management (Gianforte, 2003) via chat rooms/tools. In wed-based chat rooms and BBS a large volume of NIL text is abused by (McCullagh, 2004). A survey by the Global System for Mobile Communication (GSM) showed that Germans send 200 million messages a year (German News, 2004). All the facts disclose the growing importance in processing NIL text.

Chat text holds anomalous characteristics in forming non-alphabetical characters, words, and phrases. It uses ill-edited terms and anomalous writing styles. Typical examples of anomalous Chinese chat terms can be found in (Xia et. al., 2005a). Besides the anomalous characteristics, our observations reveal remarkable dynamic nature of the chat text. The anomaly is created and discarded very quickly. Although there is no idea how tomorrow's chat text would look like, the changing will never stop. Instead, the changing gets faster and faster.

The challenging issues originates from the dynamic nature are two-fold. On the one hand, anomalous chat terms and writing styles are frequently found in chat text. Knowledge about chat text is urgently required to understand the anomaly. On the other hand, the dynamic nature of the chat text makes it nearly impossible to maintain a timely chat text knowledge base. This claim has been proved by (Xia et. al., 2005a) in which experiments are conducted with an SVM classifier. The classifier is trained on chat text created in an earlier period and tested on chat text created in a later period. In their experiments, performance of the SVM classifier becomes lower when the two periods are farther. This reveals that chat text is written in such a style that changes constantly along with time. A straightforward solution to this problem is to re-train the SVM classifier periodically with timely chat text collections. Unfortunately, this solution costs a lot of manpower in producing new chat text corpora. The super-

vised learning technique becomes ineffective in processing chat text.

This paper proposes approaches to detecting anomaly in dynamic Chinese chat text by incorporating standard Chinese corpora and a static chat corpus. The idea is basically error-driven. That is, we first create standard language models using trigram on standard Chinese corpora. These corpora provide negative training samples. We then construct anomalous chat language model using one static chat text corpus which provides positive training samples. We incorporate the chat language model with the standard language models and calculate confidence and entropy to help make decisions whether input text is anomalous chat text. We investigate two types of trigram, i.e. word trigram and part-of-speech (POS) tag trigram in this work.

The remaining sections of this paper are organized as follow. In Section 2, the works related to this paper are addressed. In Section 3, approaches of anomaly detection in dynamic Chinese chat text with standard Chinese corpora are presented. In Section 4, we incorporate the NIL corpus into our approaches. In section 5, experiments are described to estimate threshold values and to evaluate performance of the two approaches with various configurations. Comparisons and discussions are also reported. We conclude this paper and address future works in Section 6.

2 Related Works

Some works had been carried out in (Xia et. al., 2005a) in which an SVM classifier is implemented to recognize anomalous chat text terms. A within-domain open test is conducted on chat text posted in March 2005. The SVM classifier is trained on five training sets which contain chat text posted from December 2004 to February 2005. The experiments show that performance of the SVM classifier increases when the training period and test period are closer. This reveals that chat text is written in a style that changes quickly with time. Many anomalous popular chat terms in last year are forgotten today and new ones replace them. This makes SVM based pattern learning technique ineffective to reflect the changes.

The solution to this problem in (Xia et. al., 2005b) is to re-train the SVM classifier periodically. This costs a lot of manpower in producing the timely chat text corpora, in which each piece of anomalous chat text should be annotated with several attributes manually.

We argue that the anomalous chat text can be identified using negative training samples in static Chinese corpora. Our proposal is that we model the standard natural language using standard Chinese corpora. We incorporate a static chat text corpus to provide positive training samples to reflect fundamental characteristics of anomalous chat text. We then apply the models to detect the anomalous chat text by calculating confidence and entropy.

Regarding the approaches proposed in this paper, our arguments are, 1) the approaches can achieve performance equivalent to the best ones produced by the approaches in existence; and 2) the good performance can be achieved stably. We prove these arguments in the following sections.

3 Anomaly Detection with Standard Chinese Corpora

Chat text exhibits anomalous characteristics in using or forming words. We argue that the anomalous chat text, which is referred as anomaly in this article, can be identified with language models constructed on standard Chinese corpora with some statistical language modeling (SLM) techniques, e.g. trigram model.

The problem of anomaly detection can be addressed as follows. Given a piece of anomalous chat text, i.e. $W = \{w_1, w_2, ..., w_n\}$, and a language model $LM = \{p(x)\}$, we attempt to recognize W as anomaly by the language model. We propose two approaches to tackle this problem. We design a confidence-based approach to calculate how likely that W fits into the language model. Another approach is designed based on entropy calculation. Entropy method was originally proposed to estimate how good a language model is. In our work we apply this method to estimate how much the constructed language models are able to reflect the corpora properly based on the assumption that the corpora are sound and complete.

Although there exist numerous statistical methods to construct a natural language model, the objective of them is one: to construct a probabilistic distribution model $p(x)$ which fits to the most extent into the observed language data in the corpus. We implement the trigram model and create language models with three Chinese corpora, i.e. People's Daily corpus, Chinese Gigaword and Chinese Pen Treebank. We investigate

quality of the language models produced with these corpora.

3.1 The N-gram Language Models

N-gram model is the most widely used in statistical language modeling nowadays. Without loss of generality we express the probability of a word sequence $W = \{w_1,...,w_n\}$ of *n* words, i.e. $p(W)$ as

$$p(W) = p(w_1,...,w_n) = \prod_{i=1}^{n} p(w_i \mid w_0, w_1,..., w_{i-1}) \quad (1)$$

where w_0 is chosen appropriately to handle the initial condition. The probability of the next word w_i depends on the history h_i of words that have been given so far. With this factorization the complexity of the model grows exponentially with the length of the history.

One of the most successful models of the past two decades is the trigram model (n=3) where only the most recent two words of the history are used to condition the probability of the next word.

Instead of using the actual words, one can use a set of word classes. Classes based on the POS tags, or the morphological analysis of words, or the semantic information have been tried. Also, automatically derived classes based on some statistical models of co-occurrence have been tried (Brown et. al., 1990). The class model can be generally described as

$$p(W) = \prod_{i=1}^{n} p(w_i \mid c_i)\, p(c_i \mid c_{i-2}, c_{i-1}) \quad (2)$$

if the classes are non-overlapping. These tri-class models have had higher perplexities than the corresponding trigram model. However, they have led to a reduction in perplexity when linearly combined with the trigram model.

3.2 The Confidence-based Approach

Given a piece of chat text $W = \{w_1, w_2,..., w_n\}$ where each word w_i is obtained with a standard Chinese word segmentation tool, e.g. ICTCLAS. As ICTCLAS is a segmentation tool based on standard vocabulary, it means that some unknown chat terms (e.g., "介个") would be broken into several element Chinese words (i.e., "介" and "个" in the above case). This does not hurt the algorithm because we use trigram in this method. A chat term may produce some anomalous word trigrams which are evidences for anomaly detection.

We use non-zero probability for each trigram in this calculation. This is very simple but naïve. The calculation seeks to produce a so-called confidence, which reflects how much the given text fits into the training corpus in arranging its element Chinese words. This is enlightened by the observation that the chat terms use element words in anomalous manners which can not be simulated by the training corpus.

The confidence-based value is defined as

$$C(W) = \left(\prod_{i=1}^{K} C(T_i) \right)^{\frac{1}{K}} \quad (3)$$

where *K* denotes the number of trigrams in chat text *W* and T_i is the *i*-th order trigram. $C(T_i)$ is confidence of trigram T_i. Generally $C(T_i)$ is assigned probability of the trigram T_i in training corpus, i.e. $p(T_i)$. When a trigram is missing, linear interpolation is applied to estimate its probability.

We empirically setup a confidence threshold value to determine whether the input text contains chat terms, namely, it is a piece of chat text. The *input* is concluded to be *stand* text if its confidence is bigger than the confidence threshold value. Otherwise, the *input* is concluded to be *chat* text. The confidence threshold value can be estimated with a training chat text collection.

3.3 The Entropy-based Approach

The idea beneath this approach comes from entropy based language modeling. Given a language model, one can use the quantity of entropy to get an estimation of how good the language model (LM) might be. Denote by *p* the true distribution, which is unknown to us, of a segment of new text *x* of *k* words. Then the entropy on a per word basis is defined as

$$H = \lim_{k->\infty} -\frac{1}{k} \sum_{x} p(x) \ln p(x) \quad (4)$$

If every word in a vocabulary of size |*V*| is equally likely then the entropy would be $\log_2 |V|$; $H \le \ln |V|$ for other distributions of the words.

Enlightened by the estimation method, we compute the entropy-based value on a per trigram basis for the input chat text. Given a standard LM denoted by $\tilde{p}$ which is modeled by trigram, the entropy-value is calculate as

$$\tilde{H}_K = -\frac{1}{K}\sum_{i=1}^{K}\tilde{p}(T_i)\ln\tilde{p}(T_i) \quad (5)$$

where K denotes number of trigrams the input text contains. Our goal is to find how much difference the input text is compared against the LM. Obviously, bigger entropy discloses a piece of more anomalous chat text. An empirical entropy threshold is again estimated on a training chat text collection. The *input* is concluded to be *stand* text if its entropy is smaller than the entropy threshold value. Otherwise, the *input* is concluded to be *chat* text.

4 Incorporating the Chat Text Corpus

We argue performance of the approaches can be improved when an initial static chat text corpus is incorporated. The chat text corpus provides some basic forms of the anomalous chat text. These forms we observe provide valuable heuristics in the trigram models. Within the chat text corpus, we only consider the word trigrams and POS tag trigrams in which anomalous chat text appears. We thus construct two trigram lists. Probabilities are produced for each trigram according to its occurrence. One chat text example EXP1 is given below.

EXP1: 介个故事听起来 8 错。
SEG1: 介 个 故事 听 起来 8 错 。

SEG1 presents the word segments produced by ICTCLAS. We generate chat text word trigrams based on SEG1 as follow.

TRIGRAM1: (1)/介 个 故事/
(2)/个 起来 8 /
(3)/起来 8 错/
(4)/ 8 错 。/

For each input trigram T_i, if it appears in the chat text corpus, we adjust the confidence and entropy values by incorporating its probability in chat text corpus.

4.1 The Refined Confidence

For each $C(T_i)$, we assign a weight ϖ_i, which is calculated as

$$\varpi_i = e^{p_n(T_i)-p_c(T_i)} \quad (6)$$

where $p_n(T_i)$ is probability of the trigram T_i in standard corpus and $p_c(T_i)$ probability in chat text corpus. Equation (3) therefore is re-written as

$$\begin{aligned} C'(W) &= \left(\prod_{i=1}^{K}\varpi_i C(T_i)\right)^{\frac{1}{K}} \\ &= \left(\prod_{i=1}^{K} e^{p_n(T_i)-p_c(T_i)}\, p_n(T_i)\right)^{\frac{1}{K}} \end{aligned} \quad (7)$$

The intention of inserting ϖ_i into confidence calculation is to decrease confidence of input chat text when chat text trigrams are found. Normally, when a trigram T_i is found in chat text trigram lists, $p_n(T_i)$ will be much lower than $p_c(T_i)$; therefore ϖ_i will be much lower than 1. By multiplying such a weight, confidence of input chat text can be decreased so that the text can be easily detected.

4.2 The Refined Entropy

Instead of assigning a weight, we introduce the entropy-based value of the input chat text on the chat text corpus, i.e. $\tilde{H}_K^c$, to produce a new equation. We denote $\tilde{H}_K^n$ the entropy calculated with equation (5). Similar to $\tilde{H}_K^n$, $\tilde{H}_K^c$ is calculated with equation (8).

$$\tilde{H}_K^c = -\frac{1}{K}\sum_{i=1}^{K}\tilde{p}_c(T_i)\ln\tilde{p}_c(T_i) \quad (8)$$

We therefore re-write the entropy-based value calculation as follows.

$$\begin{aligned} \tilde{H}_K &= \tilde{H}_K^n + \tilde{H}_K^c \\ &= -\frac{1}{K}\sum_{i=1}^{K}\left(\tilde{p}_n(T_i)\ln\tilde{p}_n(T_i) + \tilde{p}_c(T_i)\ln\tilde{p}_c(T_i)\right) \end{aligned} \quad (9)$$

The intention of introducing $\tilde{H}_K^c$ in entropy calculation is to increase the entropy of input chat text when chat text trigrams are found. It can be easily proved that $\tilde{H}_K$ is never smaller than H_K^n. As bigger entropy discloses a piece of more anomalous chat text, we believe more anomalous chat texts can be correctly detected with equation (9).

5 Evaluations

Three experiments are conducted in this work. The first experiment aims to estimate threshold values from a real text collection. The remaining experiments seek to evaluate performance of the approaches with various configurations.

5.1 Data Description

We use two types of text corpora to train our approaches in the experiments. The first type is

standard Chinese corpus which is used to construct standard language models. We use People's Daily corpus, also know as Peking University Corpus (PKU), the Chinese Gigaword (CNGIGA) and the Chinese Penn Treebank (CNTB) in this work. Considering coverage, CNGIGA is the most excellent one. However, PKU and CPT provide more syntactic information in their annotations. Another type of training corpus is chat text corpus. We use NIL corpus described in (Xia et. al., 2005b). In NIL corpus each anomalous chat text is annotated with their attributes.

We create four test sets in our experiments. We use the test set #1 to estimate the threshold values of confidence and entropy for our approaches. The values are estimated on two types of trigrams in three corpora. Test set #1 contains 89 pieces of typical Chinese chat text selected from the NIL corpus and 49 pieces of standard Chinese sentences selected from online Chinese news by hand. There is no special consideration that we select different number of chat texts and standard sentences in this test set.

The remaining three test sets are used to compare performance of our approaches on test data created in different time periods. The test set #2 is the earliest one and #4 the latest one according to their time stamp. There are 10K sentences in total in test set #2, #3 and #4. In this collection, chat texts are selected from YESKY BBS system (http://bbs.yesky.com/bbs/) which cover BBS text in March and April 2005 (later than the chat text in the NIL corpus), and standard texts are extracted from online Chinese news randomly. We describe the four test sets in Table 1.

Test set	# of standard sentences	# of chat sentences
#1	49	89
#2	1013	2320
#3	1013	2320
#4	1014	2320

Table 1: Number of sentences in the four test sets.

5.2 Experiment I: Threshold Values Estimation

5.2.1 Experiment Description

This experiment seeks to estimate the threshold values of confidence and entropy for two types of trigrams in three Chinese corpora.

We first run the two approaches using only standard Chinese corpora on the 138 sentences in the first test set. We put the calculated values (confidence or entropy) into two arrays. Note that we already know type of each sentence in the first test set. So we are able to select in each array a value that produces the lowest error rate. In this way we obtain the first group of threshold values for our approaches.

We incorporate the NIL corpus to the two approaches and run them again. We then produce the second group of threshold values in the same way to produce the first group of values.

5.2.2 Results

The selected threshold values and corresponding error rates are presented in Table 2~5.

Trigram option	Threshold	Err rate
word of CNGIGA	1.58E-07	0.092
word of PKU	7.06E-07	0.098
word of CNTB	2.09E-06	0.085
POS tag of CNGIGA	0.0278	0.248
POS tag of PKU	0.0143	0.263
POS tag of CNTB	0.0235	0.255

Table 2: Selected threshold values of confidence for the approach using standard Chinese corpora and error rates.

Trigram option	Threshold	Err rate
word of CNGIGA	3.762E-056	0.099
word of PKU	5.683E-048	0.112
word of CNTB	2.167E-037	0.169
POS tag of CNGIGA	0.00295	0.234
POS tag of PKU	0.00150	0.253
POS tag of CNTB	0.00239	0.299

Table 3: Selected threshold values of entropy for the approach using standard Chinese corpora and error rates.

Trigram option	Threshold	Err rate
word of CNGIGA	4.26E-05	0.089
word of PKU	3.75E-05	0.102
word of CNTB	6.85E-05	0.092
POS tag of CNGIGA	0.0398	0.257
POS tag of PKU	0.0354	0.266
POS tag of CNTB	0.0451	0.249

Table 4: Selected threshold values of confidence for the approach incorporating the NIL corpus and error rates.

Trigram option	Threshold	Err rate
word of CNGIGA	8.368E-027	0.102
word of PKU	3.134E-019	0.096
word of CNTB	5.528E-021	0.172
POS tag of CNGIGA	0.00465	0.241
POS tag of PKU	0.00341	0.251
POS tag of CNTB	0.00532	0.282

Table 5: Selected thresholds values of entropy for the approach incorporating the NIL corpus and error rates.

We use the selected threshold values in experiment II and III to detect anomalous chat text within test set #2, #3 and #4.

5.3 Experiment II: Anomaly Detection with Three Standard Chinese Corpora

5.3.1 Experiment Description

In this experiment, we run the two approaches using the standard Chinese corpora on test set #2. The threshold values estimated in experiment I are applied to help make decisions.

Input text can be detected as either standard text or chat text. But we are only interested in how correctly the anomalous chat text is detected. Thus we calculate precision (p), recall (r) and F_1 measure (f) only for chat text.

$$p = \frac{a}{a+c} \quad r = \frac{a}{a+b} \quad f = \frac{2 \times p \times r}{p+r} \qquad (10)$$

where a is the number of true positives, b the false negatives and c the false positives.

5.3.2 Results

The experiment results for the approaches using the standard Chinese corpora on test set #2 are presented in Table 6.

5.3.3 Discussions

Table 4 shows that, in most cases, the entropy-based approach outperforms the confidence-based approach slightly. It can thus be conclude that the entropy-based approach is more effective in anomaly detection.

It is also revealed that both approaches perform better with word trigrams than that with POS tag trigrams. This is natural for class based trigram model when number of class is small. Thirty-nine classes are used in ICTCLAS in POS tagging Chinese words.

When the three Chinese corpora are compared, the CNGIGA performs best in the confidence-based approach with word trigram model. However, it is not the case with POS tag trigram model. Results of two approaches on CNTB are best amongst the three corpora. Although we are able to draw the conclusion that bigger corpora yields better performance with word trigram, the same conclusion, however, does not work for POS tag trigram. This is very interesting. The reason we can address on this issue is that CNTB probably provides highest quality POS tag trigrams and other corpora contain more noisy POS tag trigrams, which eventually decreases the performance. An observation on word/POS tag lists for three Chinese corpora verifies such a claim. Text in CNTB is best-edited amongst the three.

5.4 Experiment III: Anomaly Detection with NIL Corpus Incorporated

5.4.1 Experiment Description

In this experiment, we incorporate one chat text corpus, i.e. NIL corpus, to the two approaches. We run them on test set #2, #3 and #4 with the estimated threshold values. We use precision, recall and F_1 measure again to evaluate performance of the two approaches.

5.4.2 Results

The experiment results are presented in Table 7~ Table 9 on test set #2, #3 and #4 respectively.

5.4.3 Discussions

We first compare the two approaches with different running configurations. All conclusions made in experiment II still work for experiment III. They are, i) the entropy-based approach outperforms the confidence-based approach slightly in most cases; ii) both approach perform better with word trigram than POS tag trigram; iii) both approaches perform best on CNGIGA with word trigram model. But with POS tag trigram model, CNTB produces the best results.

An interesting comparison is conducted on F_1 measure between the approaches in experiment II and experiment III on test set #2 in Figure 1 (the left two columns). Generally, F_1 measure of anomaly detection with both approaches with word trigram model is improved when the NIL corpus is incorporated. It is revealed in Table 7~9 that same observation is found with POS tag trigram model.

We compare F_1 measure of the approaches with word trigram model in experiment III on test set #2, #3 and #4 in Figure 1 (the right three columns). The graph in Figure 1 shows that F_1 measure on three test sets are very close to each other. This is also true the approaches with POS tag trigram model as showed in Table 7~9. This provides evidences for the argument that the approaches can produce stable performance with the NIL corpus. Differently, as reported in (Xia et. al., 2005a), performance achieved in SVM classifier is rather unstable. It performs poorly with training set C#1 which contains BBS text posted several months ago, but much better with training set C#5 which contains the latest chat text.

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.685	0.737	0.710	0.722	0.761	0.741	0.614	0.654	0.633	0.637	0.664	0.650
PKU	0.699	0.712	0.705	0.701	0.738	0.719	0.619	0.630	0.624	0.625	0.648	0.636
CNTB	0.653	0.661	0.657	0.692	0.703	0.697	0.651	0.673	0.662	0.684	0.679	0.681

Table 6: Results of anomaly detection using standard Chinese corpora on test set #2.

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.821	0.836	0.828	0.857	0.849	0.853	0.653	0.657	0.655	0.672	0.678	0.675
PKU	0.818	0.821	0.819	0.838	0.839	0.838	0.672	0.672	0.672	0.688	0.679	0.683
CNTB	0.791	0.787	0.789	0.821	0.811	0.816	0.691	0.679	0.685	0.712	0.688	0.700

Table 7: Results of anomaly detection incorporating NIL corpus on test set #2

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.819	0.841	0.830	0.849	0.848	0.848	0.657	0.659	0.658	0.671	0.677	0.674
PKU	0.812	0.822	0.817	0.835	0.835	0.835	0.663	0.671	0.667	0.687	0.681	0.684
CNTB	0.801	0.783	0.792	0.822	0.803	0.812	0.689	0.677	0.683	0.717	0.689	0.703

Table 8: Results of anomaly detection incorporating NIL corpus on test set #3

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.824	0.839	0.831	0.852	0.845	0.848	0.651	0.654	0.652	0.674	0.674	0.674
PKU	0.815	0.825	0.820	0.836	0.84	0.838	0.668	0.668	0.668	0.692	0.682	0.687
CNTB	0.796	0.785	0.790	0.817	0.807	0.812	0.694	0.681	0.687	0.713	0.686	0.699

Table 9: Results of anomaly detection incorporating NIL corpus on test set #4

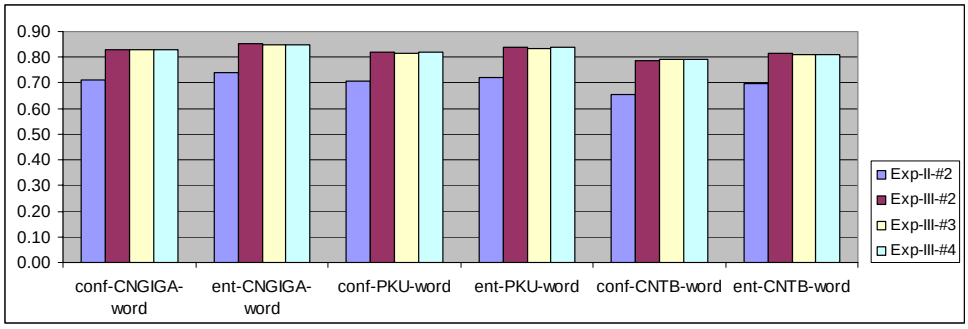

Figure 1: Comparisons on F_1 measure of the approaches with word trigram on test set #2, #3 and #4 in experiment II and experiment III.

We finally compare performance of our approaches against the one described in (Xia, et. al., 2005a). The best F_1 measure achieved in our work, i.e. 0. 853, is close to the best one in their work, i.e. 0.871 with training corpus C#5. This proves another argument that our approaches can produce equivalent performance to the best ones achieved by the approaches in existence.

6 Conclusions

The new approaches to detecting anomalous Chinese chat text are proposed in this paper. The approaches calculate confidence and entropy values with the language models constructed on negative training samples in three standard Chi-

nese corpora. To improve detection quality, we incorporate positive training samples in NIL corpus in our approaches. Two conclusions can be made based on this work. Firstly, F_1 measure of anomaly detection can be improved by around 0.10 when NIL corpus is incorporated into the approaches. Secondly, performance equivalent to the best ones produced by the approaches in existence can be achieved stably by incorporating the standard Chinese corpora and the NIL corpus.

We believe some strong evidences for our claims can be obtained by training our approaches with more chat text corpora which contain chat text created in different time periods. We are conducting this experiment seeks to find out whether and how our approaches are independent of time. This work is still progressing. A report on this issue will be available shortly. We also plan to investigate how size of chat text corpus influences performance of our approaches. The goal is to find the optimal size of chat text corpus which can achieve the best performance. The readers should also be noted that evaluation in this work is a within-domain test. Due to shortage of chat text resources, no cross-domain test is conducted. In the future cross-domain test, we will investigate how our approaches are independent of domain.

Eventual goal of chat text processing is to normalize the anomalous chat text, namely, convert it to standard text holding the same meaning. So the work carried out in this paper is the first step leading to this goal. Approaches will be designed to locate the anomalous terms in chat text and map them to standard words.

Acknowledgement

Research described in this paper is partially supported by the Chinese University of Hong Kong under the Direct Grant Scheme (No: 2050330) and Strategic Grant Scheme project (No: 4410001) respectively.

Reference

Brown, P. F., V. J. Della Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer. 1990. Class-based n-gram models of natural language. In Proceedings of the IBM Natural Language ITL, Paris, France.

Finkelhor, D., K. J. Mitchell, and J. Wolak. 2000. Online Victimization: A Report on the Nation's Youth. Alexandria, Virginia: National Center for Missing & Ex-ploited Children, page ix.

German News. 2004. Germans are world SMS champions, 8 April 2004, http://www.expatica.com/source/site_article.asp?subchannel_id=52&story_id=6469.

Gianforte, G.. 2003. From Call Center to Contact Center: How to Successfully Blend Phone, Email, Web and Chat to Deliver Great Service and Slash Costs. RightNow Technologies.

Heard-White, M., Gunter Saunders and Anita Pincas. 2004. Report into the use of CHAT in education. Final report for project of Effective use of CHAT in Online Learning, Institute of Education, University of London.

McCullagh, D.. 2004. Security officials to spy on chat rooms. News provided by CNET Networks. November 24, 2004.

Xia, Y., K.-F. Wong and W. Gao. 2005a. NIL is not Nothing: Recognition of Chinese Network Informal Language Expressions, 4th SIGHAN Workshop on Chinese Language Processing at IJCNLP'05, pp95-102.

Xia, Y., K.-F. Wong and R. Luk. 2005b. A Two-Stage Incremental Annotation Approach to Constructing A Network Informal Language Corpus. In Proc. of NTCIR-5 Meeting, pp. 529-536.

Zhang, Z., H. Yu, D. Xiong and Q. Liu. 2003. HMM-based Chinese Lexical Analyzer ICTCLAS. SIGHAN'03 within ACL'03, pp. 184-187.