Observing Features of PTT Neologisms:

A Corpus-driven Study with N-gram Model

Tsun-Jui Liu
Graduate Institute of Linguistics
National Taiwan University
r99142008@ntu.edu.tw

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Laurent PREVOT
Laboratoire Parole et Langage
Université Aix-Marseille
laurent.prevot@lpl-aix.fr

Abstract

PTT (批踢踢) is one of the largest web forums in Taiwan. In the last few years, its importance has been growing rapidly because it has been widely mentioned by most of the mainstream media. It is observed that its influence reflects not only on the society but also on the language novel use in Taiwan. In this research, a pipeline processing system in Python was developed to collect the data from PTT, and the n-gram model with proposed linguistic filter are adopted with the attempt to capture two-character neologisms emerged in PTT. Evaluation task with 25 subjects was conducted against the system's performance with the calculation of Fleiss' kappa measure. Linguistic discussion as well as the comparison with time series analysis of frequency data are provided. It is hoped that the detection of neologisms in PTT can be improved by observing the features, which may even facilitate the prediction of the neologisms in the future.

Keywords: **PTT, Neologisms, n-gram, Fleiss' kappa, Time series analysis**

1. Introduction

A neologism in general refers to "a newly coined term, word, or phrase, that may be in the process of entering common use, but has not yet been accepted into mainstream language" (Levchenko, 2010)[1]. It is closely related to the *unknown words* or *out-of-vocabulary* in the field of Speech and Natural Language Processing, but with the nuance that the latter is often formally defined by its non-existence in a given vocabulary repository. With the emergence of voluminous data on the web and fast-developing technologies, never before has our world been facing with such an overwhelming mass of neologisms. Therefore, the description and

[1] As cited by wiki at http://en.wikipedia.org/wiki/Neologism#cite_ref-1

detection of neologism has become an important research topic in the recent years.

In this paper, we aim to begin with a corpus-driven approach in exploring the linguistic features of Chinese neologisms. We use PTT as our corpus data. As widely known, PTT is one of the largest web forums in Taiwan that contain users from various backgrounds and ages. In these years, its importance has been growing rapidly because it has been widely mentioned by most of the mainstream media in Taiwan. As Magistry (2012) suggested, "PTT should be seen as an extension of the modern society in Taiwan." This implies that PTT has great influence not only on the society but also the novel language use in Taiwan, which motives this research to exploit PTT as data source.

Section 2 explains the pipeline framework developed for data crawling and pre-processing, and the lexicon and filter for capturing two-character neologisms in PTT. Section 3 introduces the methodological part, where the rationale of our proposed 'diachronic n-gram model' is introduced and classification results are shown. Section 4 provides the discussion on the evaluation task as well as explanation from linguistic perspective. A time series analysis on the extracted diachronic n-gram data is conducted for further investigation. Section 5 concludes this paper.

2. Corpus Data

2.1. PTT

PTT (批踢踢)[2], founded in 1999, is a terminal-based bulletin board system (BBS) based in Taiwan. It is a non-profit, free and open online community, and it is claimed to be one of the largest BBS sites in the world. PTT contains over 20,000 discussion boards with more than 1.5 million registered users, and over 10,000 articles are posted every day. The screenshot of PTT is shown in Figure 1.

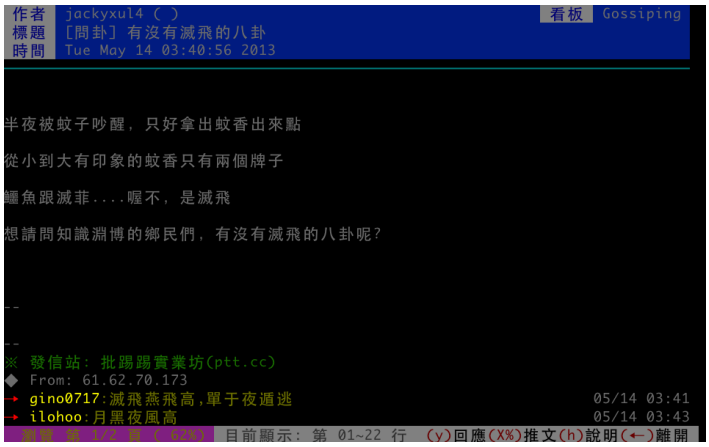

Figure 1. Screenshot of PTT

The data are collected from 2005 to 2012 from three major boards on PTT, which are Gossiping (八卦版), joke (就可版) and StupidClown (笨版). Figure 2 shows the number of tokens in the corpus per year, and Table 1 provides some basic meta-information

[2] telnet://ptt.cc

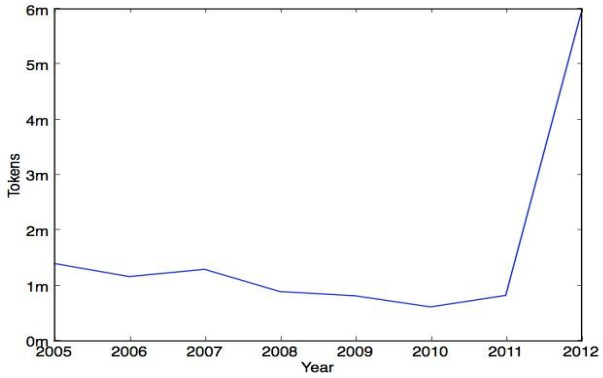

Figure 2. Number of tokens in the corpus per year from 2005 to 2012

Table 1. PTT Corpus

Boards	Gossiping(八卦版), joke (就可版), StupidClown (笨版)
Years	2005 – 2012
Posts	33,450
Authors	17,031
Tokens	14,285,768
Types	7,010
Bigrams	785,494

2.2. Lexicon

In this research, the lexicon was used for filtering out existed words. It is comprised of The Revised Chinese Dictionary (教育部重編國語辭典修訂本, TRCD)[3] and Taiwan Spoken Mandarin Wordlist (中研院漢語口語語料庫詞頻表, TSMW)[4]. TRCD was compiled by Ministry of Education with 139,401 words and expressions, and TSMW was collected by Academia Sinica with 16,683 entries. Since two-character words are dominant in modern Chinese, as a first step, only two-character words will be chosen.

Table 2. Lexicon

	TRCD	TSMW
Entries	159,401	16,683
Two-character word	86,907	10,198
Two-character words in total: 89,118		

2.3. Data pre-processing

We have developed a pipeline framework for the corpus-driven analysis. A crawler module

[3] http://dict.revised.moe.edu.tw/
[4] http://mmc.sinica.edu.tw/resources_c_01.htm

collects the textual data and meta-information from the PTT; a cleaner module removes the unnecessary information of the retrieved raw data; an n-gram module creates bigram candidates and compares them with the lexicon; and finally a linguistic module filters out some noisy data via encodes heuristic rules[5]. The resulting bigrams are thus divided into three basic categories: *words*, *nonwords* and *potential neologisms*. The main steps can be listed as follows and illustrated in Figure 3:

Step 1. Transforming all the tokens into bigrams
Step 2. Exploiting the lexicon to exclude existed words from out-of-vocabulary (OOV)
Step 3. Linguistics rules were applied to separate OOV into nonwords and potential neologisms

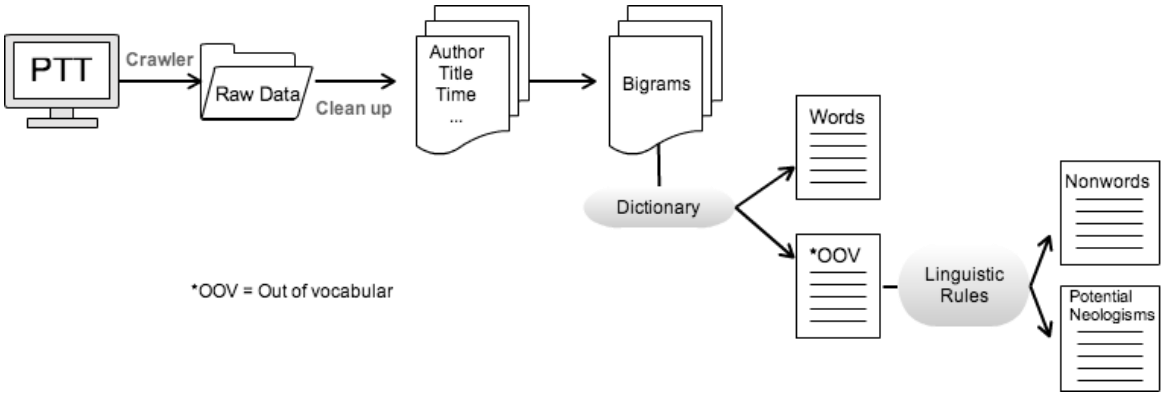

Figure 3. Data processing flowchart

3. Methodology

Most previous works on unknown word / OOV extraction exploited complicated morphological rules and various machine learning techniques (Chen and Ma, 2002). In order to utilize the contextual information, as much linguistic resource (such as syntax, semantics, morphology and world knowledge) as possible were explored. It is worth mentioning that why the (naive) n-gram model is adopted in this study.

3.1. *N*-gram in Diachronic Contexts

An *n*-gram is a contiguous sequence of *n* items from a given sequence of text. In Mandarin Chinese, the items correspond to individual characters. The *n*-grams of size 2, viz. *bigrams*, will be the major focus in this research. A bigram is a sequence of two adjacent elements in a

[5] Linguistic rules are used to exclude bigrams with function words or affixes, such as pronouns, particles and aspects. See Li and Thompson (1989).

string of tokens. For example, there are five bigrams in *今天天氣很好*, which are *今天*, *天天*, *天氣*, *氣很* and *很好*. In this paper, we further propose a notion of 'diachronic n-gram' by leveraging diachronic frequency data in PTT, whose advantages can be explicated by the following points:

First, this model does not have to presume a word segmentator. The reason why prominent segmentation system such as CKIP[8] was not used to segmentate words is that language used on PTT contains too many fragments, novel linguistic forms, jargons and slangs, causing the low accuracy of the performance. Take the following sentences as an example. Sentence (1) is a sentence extracted from the data, and sentence (2) is the segmentation result by CKIP.

(1) 小妹想請問各位批踢踢帥宅宅葛格們

(2) 小妹/想/請問/各/位/批/踢踢/帥宅宅葛格們

As we can see, the result of segmentation is out of satisfactory. Stenetorp (2010) suggested "[…] an exclusion error is not recoverable and likely to make users unable to observe a certain neologism we might be forced to tolerate a high degree of noise." To reduce the risk of losing any potential neologism, segmentator was not exploited in this research.

Secondly, an n-gram model equipped with diachronic information would arouse echoes in current theoretical development in linguistics. *Frequency effect* has been widely recognized in cognitive linguistics, and recent functional linguistic studies also justify the frequency as a determinant in lexical diffusion and changes (Bybee, 2007). A usage-based perspective on language also argues that language as a complex adaptive system is to be viewed as emergent from the repeated application of underlying process, rather than given a priori or by design (Hopper, 1987). Instead of rule-based normalization, modeling lexical change with empirical data support could also bypass the thorny *wordhood* issue in Chinese. In addition, time series statistical analysis and other distributional models can bring their contribution in this scenario too.

3.2. Classification

Based on the considerations and framework mentioned above, the data was categorized into words, nonwords and potential neologisms, whose frequency data are plotted as in Figure 4.

In Figure 4, x-axis represents different time periods, which starts from 2005 to 2012, and y-axis represents the frequency of bigrams. Each curves stands for an individual bigram, and the total number of the bigrams are listed in the upper-left corner of each plot. (For example,

[8] Chinese Knowledge and Information Processing (CKIP) is a Chinese word segmentation system developed by Academia Sinica.

there are 2,836 bigrams in the category *words*.)

Generally, a first look at the data shows that the overall frequency of *words* is higher than *nonwords* and *potential neologisms*, and the frequency of *potential neologisms* is slightly lower than *nonwords*.

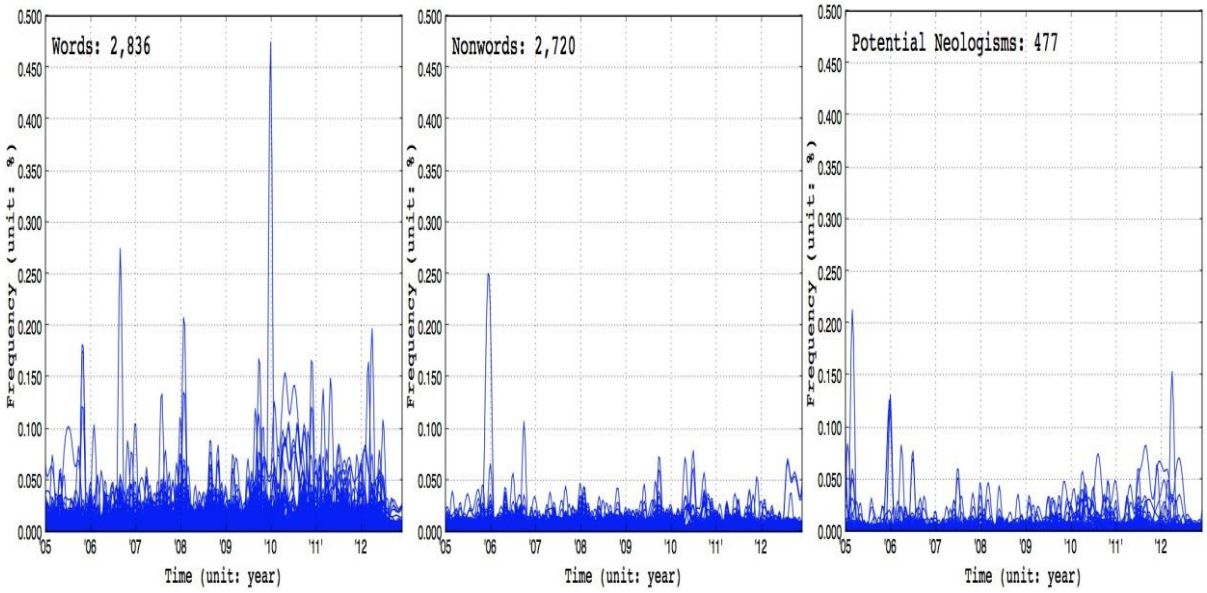

Figure 4. Plots of words, nonwords and potential neologisms

4. Evaluation and Discussions

4.1. Human Judgment Experiment

In order to evaluate the classification performance, the results were manually annotated, and measured with Fleiss' kappa (1971), a statistical measure of inter-rater reliability. The equation is shown as the following:

$$\kappa\ (Fleiss'kappa)\ = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}$$

The score of Fleiss' kappa is $\overline{P} - \overline{P}_e$, the degree of agreement actually achieved above chance, divided by $1 - \overline{P}_e$, the degree of agreement that is attainable above chance. It can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly.

In this annotation task, 25 raters (*n*) were assigned 75 bigrams (*N*, which were selected from each category randomly and equally) into three categories (*k*, i.e., words, nonwords, potential neologisms) according to the following definitions:

(1) Words: bigrams that are stable or already exist in current language use.

(2) Nonwords: bigrams that are unstable, does not exist, or being used only by a very small subculture.

(3) Potential Neologisms: bigrams that have reached a significant audience, but probably not yet have gained lasting acceptance.

The result shows that the score of Fleiss' kappa is 0.54, which indicates “moderate agreement” (Landis and Koch, 1977).

4.2. Discussions

In this section, the characteristics of the neologisms and the inconsistency between the system’s judgment and the rates’ judgment will be discussed.

For the raters’ judgment, a bigram will be recognized as a neologism if more than half of the raters have the same agreement on it. The results are categorized under Hsu’s (1999) classification, which are shown in Table 3.

Table 3. Neologism Classification (Hsu, 1999)

Native neologisms	自刪 笨點 搞笑 筆電 科大 高鐵
Loan words	馬克
Dialectal words	唬爛 白目 豪泛
Trendy words	阿罵

First, as it can be seen, *native neologisms* are in the majority of neologisms. According to Hsu, *native neologisms* appear when there is a lexical gap, and they are born without any effect from other foreign language. Second, Min Nan provides the major source of *dialectal neologism*. This shows that Min Nan has the higher prestige in Taiwanese dialects, which is also in accordance to Hsu’s proposal. It is interesting that most of the *dialectal neologisms* seem to have negative meanings, but more evidence should be provided to support this observation, which will be included the future research. Third, abbreviation words such as *筆電* and *科大* forms the major source of *native neologisms*, which corresponds to Hsu’s proposal as well. Fourth, *阿罵* is categorized as a *trendy word* since it is a play on words. That is to say, *阿罵* [a ma4] has the same pronunciation with *阿嬤* [a ma4], which is an existed word in Taiwanese Mandarin.

As mentioned earlier, 25 bigrams were randomly selected from the category *potential neologisms.* In the result, it is observed that only parts of them are rated as neologisms, and some of the bigrams originally selected from *words* and *nonwords* are rated as neologisms as well. Table 4 shows the inconsistency between the system’s judgment and raters’ judgment. Also, the last column indicates the numbers of bigrams’ occurrences in newspaper[9], which is used to show the relationship between the public news and neologisms.

[9] Newspapers are comprised of 聯合報, 經濟日報, 民生報, 聯合晚報 and Upaper with 11,230,842 articles, which are collected by United Daily News. See http://udndata.com/ndapp/Detail.

According to the number of people with agreement, we can see that *dialectal words* tend to have higher *newness* (the degree of how new a word is), showing that *dialectal words* play an important role in the input of neologisms of Taiwanese Mandarin. Second, it is shown that the higher the *newness* of a bigram, the less frequent it will appear in the public newspapers, which reflects that the more stable a bigram is, the more it will be recognized as a formal word. For example, *白目* has the lower occurrence than *科大* in the public newspapers because it has the higher newness.

Table 4. Neologisms according to raters' judgment

Bigrams	System's judgment	Number of people with agreement	Number of occurrence in newspapers
唬爛	Potential neologisms	21	127
白目		17	787
自刪		17	37
豪泫		15	6
笨點		11	3
筆電		11	13214
科大		10	17847
搞笑	Words	15	6829
高鐵		12	19893
馬克		11	4909
阿罵	Nonwords	12	2

From the statistic perspective, time series analysis also shows the similar correspondence with our prediction. The time series of the frequency data appears is *non-seasonal*, and can be probably described by using an additive model. We use Holt Winters exponential smoothing method to make short term forecast for the 4 words in the three categories. Figure 5 shows the illustrative plots for 阿罵 (nonword), 筆電 (potential neologism), 高鐵 (word), 小鬼 (word) with parameter alpha of (0.369, 0.2328, 0.0088, 0.1933) respectively. The predictive model gives us the forecast for the year 2013 (plotted as a blue line), an 80% prediction interval for the forecast (plotted as a purple shaded area,), and the 95% prediction interval as a gray shaded area.

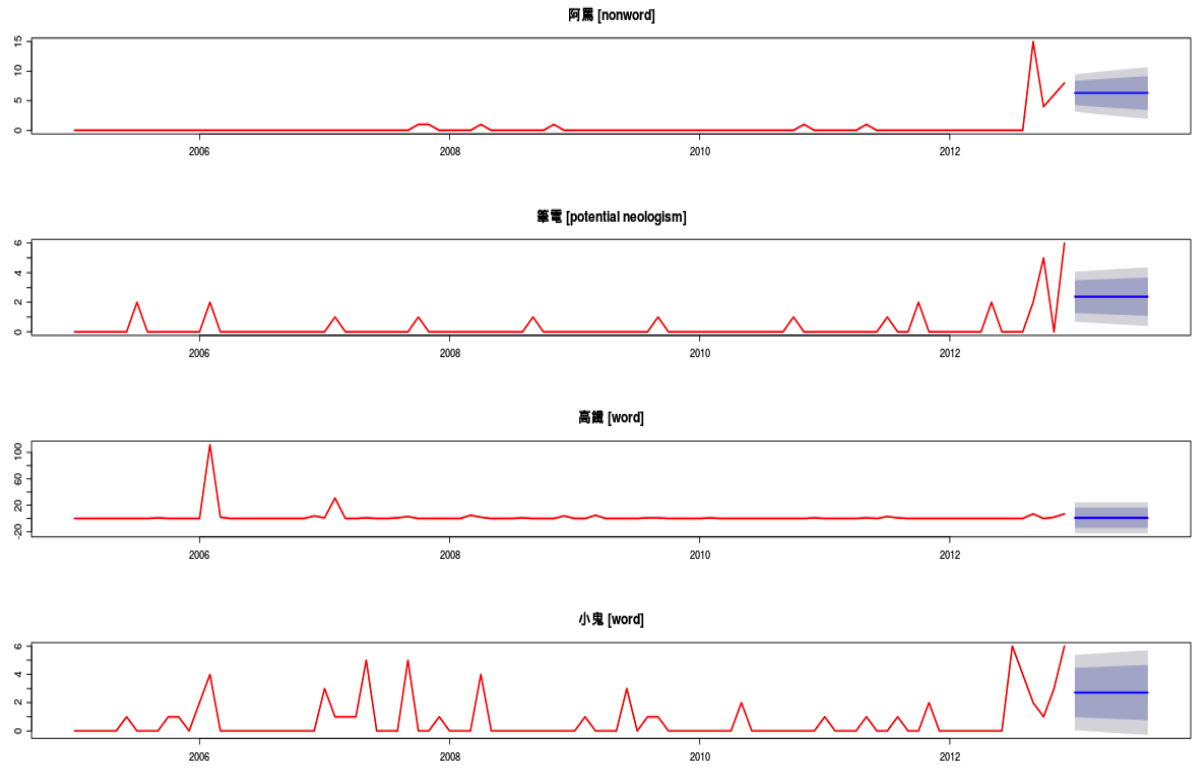

Figure 5. The time series of the frequency data

Although the overall frequency of 筆電 is low, its occurrence is relatively stable. The figure shows that it has a higher probability of being a neologism. According to this observation, we suggest that a bigram with low frequency and high stability has a higher chance of being a neologism.

In terms of the distribution, we can divide *words* into two patterns. Take *小鬼* and *高鐵* for example. The former one has peaks with high frequency during its development, which implies that it has a higher stability of being a word. The latter one has a significant peak at the beginning, and then it starts decreasing gradually. In fact, *高鐵* (Taiwan High Speed Rail) was a popular issue since late 2005 after the construction was formally announced by the government, but the topic was out of focus year after year. This reflects that public issues sometimes can dominate the occurrence of a potential neologism, and also implies that the difficulty of detecting a potential neologisms not only due to its low frequency but also due to some extralingusitics factors.

5. Conclusion

In this research, we have built a diachronic corpus of PTT from 2005 to 2012., neologisms are detected by a proposed ‘diachronic n-gram model’ inspired by functional linguistics, and an

experiment of human judgment was conducted among 25 raters. The score of the inter-rater agreement measured by Fleiss' kappa is 0.54, which indicates the moderate agreement. The characteristics of the neologisms and the inconsistency between the system's judgment and the raters' judgment are then discussed in an attempt to improve the detection of neologisms in PTT. Comparison with newly released Google book n-gram data will be conducted in the future study, which would facilitate the prediction and deeper understanding of neologisms.

References

[1] P. Magistry, "PTT 批踢踢 as a corpus," presented at the annual meeting of the European Association of Taiwan Studies, Sønderborg, 2012.

[2] K.-J. Chen and W.-Y. Ma, "Unknown word extraction for Chinese documents," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 2002, pp. 1-7.

[3] C. N. Li and S. A. Thompson, *Mandarin Chinese: A functional reference grammar*: University of California Pr, 1989.

[4] P. Stenetorp, *Automated extraction of swedish neologisms using a temporally annotated corpus*: Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan, 2010.

[5] J. Bybee, *Frequency of Use and the Organization of Language*: Oxford University Press, 2007.

[6] P. Hopper, "Emergent grammar," *Berkeley Linguistics Conference (BLS)*, vol. 13, pp. 139-157, 1987.

[7] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, vol. 76, p. 378, 1971.

[8] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159-174, 1977.

[9] 許斐絢, "台灣當代國語新詞探微," *臺灣師範大學華語文教學研究所學位論文*, 1999.