Word in a Dictionary is used by Numerous Users

Eiji Aramaki
Kyoto University/PRESTO
eiji.aramaki@gmail.com

Sachiko Maskawa
Photonic System Solutions
sachiko.maskawa@gmail.com

Mai Miyabe
Kyoto University
mai.miyabe@gmail.com

Mizuki Morita
University of Tokyo
morita.mizuki@gmail.com

Sachi Yasuda
NINJAL
yasudasac@gmail.com

Abstract

Dictionary editing requires enormous time to dis-cuss whether a word should be listed in a dictionary or not. So as to define a dictionary word, this study employs the number of word users as a novel metrics for selecting a dictionary word. In order to obtain the word user, we used about 0.25 billion tweets of approximately 100,000 people published for five months. This study compared the classification performance of various measures. The result of the experiments revealed that a word in a dictionary is used by numerous users.

1 Introduction

Dictionary editing requires numerous time to discuss whether a word should be listed in a dictionary or not. In order to define a dictionary word, this study assumes that the following two scales are essential than word frequency:

(1) Usage period: a dictionary word has been used for lager period of time.
(2) **User population:** a dictionary word has been used by more people.

The first scale is hard to measure in practice, because the usage period requires a longitudinal data. For the investigation, the second clue has more feasibility by social media resources, which enables to know the word usage for each user.

The objective of this study is to retrieve a dictionary word. This study approaches this problem by drawing on a binary classification task, which divides the words into two categories: a dictionary word (listed in a dictionary) and a out-of-dictionary word.

For the database, we have collected 0.25 billion tweets of 100,000 people from Twitter. The experimental results have revealed that a dictionary word is highly correlated with the number of word users. Although the experiment is conducted in Japanese language, the proposed method does not depend on a specified language.

2 Related Work

So far, a strong clue for dictionary editing is a word frequency. The relation between a word frequency and its coverage has been an interest for many researchers (Crowley 2003, Freeborn 2006, Burridge and Kortmann 2008). In English, the frequent 2,000 words cover 90% of spoken language (West 1953), and the most frequent 6,000 words cover 90% of written language (Francis, Kučera et al. 1982). The results in Japanese are similar to them. The frequent 10,000 words cover almost all vocabulary used in magazines (90 magazines) (NINJAL 1997), and 17,000 words cover the vocabulary spoken in television programs (Ishino 2000). Although the target media differs, they share the same findings that frequent words cover most of the corpus. This study presents another word measure.

3 Materials

This study has used two types of data: user corpus (Section 3.1), and a gold standard data (Section 3.2):

3.1 Corpus: 100,000 people tweets

This study employs Twitter as a fundamental database, because Twitter has two strong advantages for the purpose of this study: (1) it has numerous

International Joint Conference on Natural Language Processing, pages 874–877,
Nagoya, Japan, 14-18 October 2013.

users and (2) the author information is available for each tweet. This study sampled 0.25 billion tweets from 99,964 people, as described below.

- **Data collection period:** 143 days from November 3rd 2009 to March 25th 2010.
- **Number of users:** 99,964 people, as extracted based on the following three qualifications:
  - ➢ A user who posts at least 5 tweets per a month
  - ➢ Total posts contain over 5,000 words.
  - ➢ Japanese language users: at least one Japanese UTF code characters are used in the first tweet.
- **Total number of tweets:** 253,482,784 tweets (4,258,707,255 words): the words are analyzed by a morphological analyzer (Kurohashi, Nakamura et al. 1994)

3.2 Gold standard Data

The gold standard data of this study is a word listed in the *IWANAMI* Japanese Dictionary 7th edition (Nishio, Iwabuchi et al. 2009). This dictionary is one of the best selling dictionaries in Japanese.

4 Methods

The task of this study is to classify whether a word is listed in a dictionary or not. For the classification, this study employs four measures:

1. ***freq(w)***: a word frequency of a word ***w***.
2. $R_{freq}(w)$: a rank of ***freq (w)***.
3. ***user(w)***: the number of users of a word ***w***.
4. $R_{user}(w)$: a rank of ***user(w)***.

While the first two (***freq***(***w***) and $R_{freq}(w)$) are conventional measures used among the many previous researches, the other two (***user***(***w***) and $R_{user}(w)$) are newly introduced by this study.

Baseline Approach

A easy approach is to select a word which has enough frequency (more than α times). This approach is formalized as follows: $freq(w) > \alpha$.

Proposed Approach

Instead of the frequency, the proposed approach relied on the number of users (***user***(***w***)). This approach is formalized as follows: $user\ (w) > \alpha$

Another Proposed Approach

This approach makes balance between the number of users (***user***(***w***)) and the word frequency (***freq***(***w***)). If both measures stay in balance, the both ranks should equal, satisfying the following formula:

$$R_{user}(w) = R_{freq}(w)\ .$$

If a certain user prefers to use specific words, the rank of the frequency (R_{freq}) become larger than that of users (R_{users}):

$$R_{user}(w) > R_{freq}(w).$$

In the same method, a widely used word could be extracted by using the following formula:

$$R_{user}(w) < R_{freq}(w).$$

5 Experiment

5.1 Test-set: Wikipedia entry names

A test-set consists of 4,000 nouns, which are randomly sampled from Wikipedia entry names. Half of them (2,598 nouns) are listed in the dictionary (positive examples). The other 1,402 words are out-of-dictionary (negative examples).

5.2 Comparable Methods

We compared the following classification methods:

- **Rfreq:** this method selects the words whose frequency is in the top α rank: $R_{freq}(w) < \alpha$.
- **Ruser**: this method selects the words whose user size is in the top α rank: $R_{user}\ (w) < \alpha$.
- **Ruser' (weighted based):** this method is essentially based on the number of users. However, it is weighted by the frequency as follows: $-\log\ (freq(w)) \cdot user(w) < \alpha$.
- **Ruser/Rfreq**: this approach is based on the balance of two ranks: $R\text{-}Ratio < \alpha$. Here, $R\text{-}Ratio = R_{user}(w) / R_{freq}(w).$

The evaluation is conducted in possible α range (α =0～∞).

5.3 Evaluation Metrics

The methods are evaluated using information retrieval metrics:

- **Precision (P):** # of correct outputs / # of system positive outputs.
- **Recall (R):** # of correct outputs / # of positive examples (=2,598).
- **F-measure (F):** harmonic mean between the precision and the recall.

5.4 Result

The precision-recall curve for each method is presented in Figure 1. The best F-measure points of all methods are the same (Recall=1; Precision=0.6). However, the accuracies differ in the low-recall area. Basically **Ruser** (partly **Ruser/Rfreq**) showed the best performance. **Rfreq** constantly showed poor performance rather than the others. These results indicated that the number of users is an essential factor.

Figure 2 shows the distribution of dictionary words plots in $\boldsymbol{R_{freq}(w)}$ and $\boldsymbol{R_{user}(w)}$. Numerous words are distributed on the balanced line (X=Y), indicating that $\boldsymbol{R_{freq}(w)}$ and $\boldsymbol{R_{user}(w)}$ correlated with each other.

We found several outliers in the TOP-LEFT area (Y>>X), suggesting that several words have the low number of users compared to the frequency metrics. The examples of such words are presented in Table 1 (b), consisting of many out-of-dictionary words.

5.5 Discussion

This study reveals that the number of users is an important clue to classify a dictionary word. This result has a number of applications; e.g., the popular vocabulary learning, a user number-based spell checking system, and so on.

However, this study has several limitations, which comes from the following factors:

- **User bias**: Most Twitter users are 20-30 years old. This population gap might bias the results.
- **Device bias**: The type of input device, such as keyboard typing, touch pad, and input suggestion, might bias the results.
- **Twitter bias:** The length limit of Twitter (140 characters) might prefers shorter words.

Reducing the above biases is one of the remaining problems.

6 Conclusion

This study proposes a method to classify a dictionary word. We assume that a dictionary word should be used by many users. To prove this point, we have obtained the 100,000 user texts from Twitter. Then, we have evaluated various measures: a frequency based, a user based, and the ratio based. The experimental result has revealed that the number of word users is an essential indicator for classifying a dictionary word.

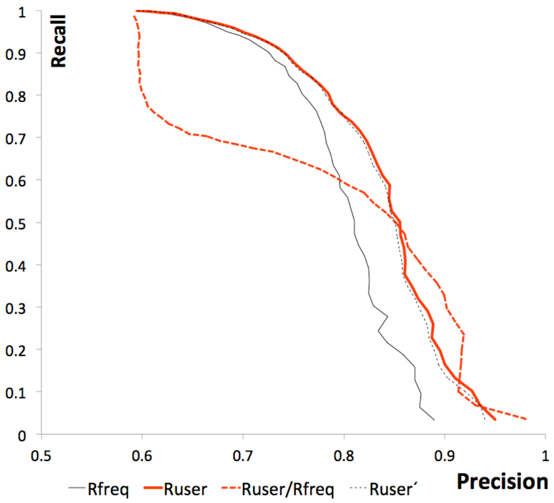


Figure 1: The precision-recall curve for each method.

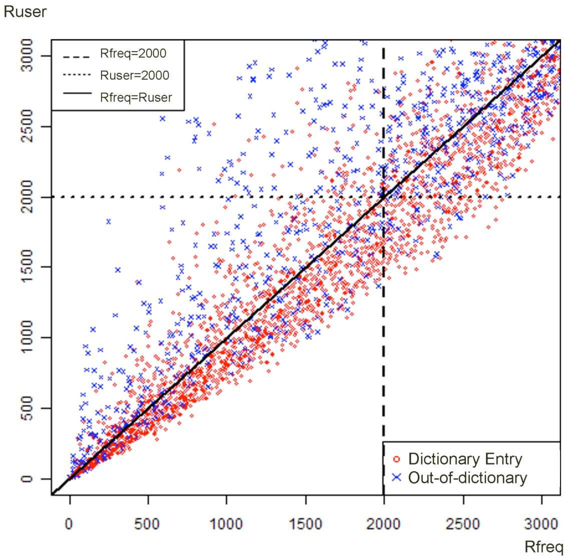
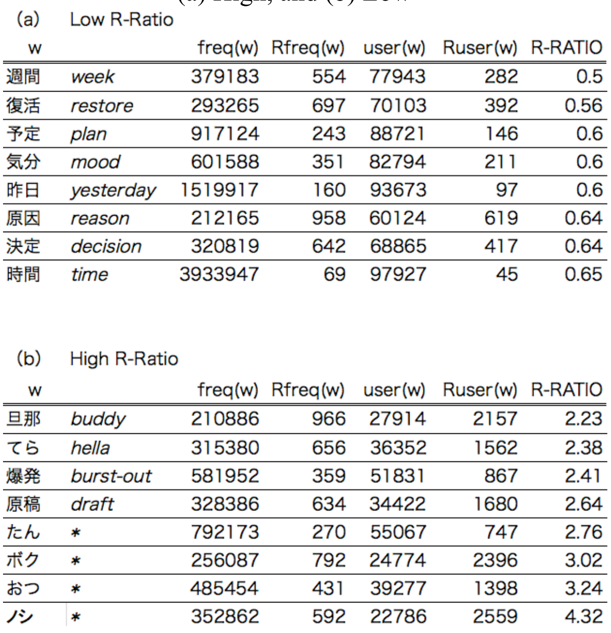


Figure 2: The Rank of Word User ($\boldsymbol{R_{user}}$) and the Rank of Word Frequency ($\boldsymbol{R_{freq}}$).

The X-axis indicates the rank of word frequency ($\boldsymbol{R_{freq}}$); the Y-axis indicates the rank of word users. The dotted line indicates $\boldsymbol{R_{freq}}$=2000 and $\boldsymbol{R_{user}}$ =2000. The line indicates that the balanced line ($\boldsymbol{R_{freq}}$= $\boldsymbol{R_{user}}$). The RIGHT-BOTTOM area contains words that are high user words. The LEFT-TOP area contains words that are low user words. As shown in the figure, most of low user words are out-of-dictionary words.

Table 1: Word Example of R_{user}/R_{freq}. (a) High, and (b) Low



(a) Low R-Ratio

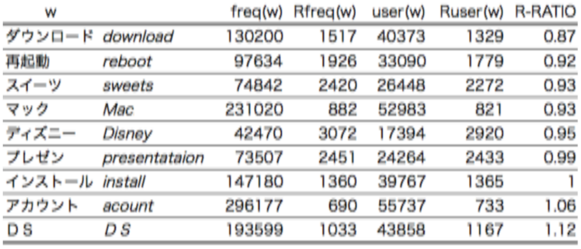
w		freq(w)	Rfreq(w)	user(w)	Ruser(w)	R-RATIO
週間	*week*	379183	554	77943	282	0.5
復活	*restore*	293265	697	70103	392	0.56
予定	*plan*	917124	243	88721	146	0.6
気分	*mood*	601588	351	82794	211	0.6
昨日	*yesterday*	1519917	160	93673	97	0.6
原因	*reason*	212165	958	60124	619	0.64
決定	*decision*	320819	642	68865	417	0.64
時間	*time*	3933947	69	97927	45	0.65

(b) High R-Ratio

w		freq(w)	Rfreq(w)	user(w)	Ruser(w)	R-RATIO
旦那	*buddy*	210886	966	27914	2157	2.23
てら	*hella*	315380	656	36352	1562	2.38
爆発	*burst-out*	581952	359	51831	867	2.41
原稿	*draft*	328386	634	34422	1680	2.64
たん	*	792173	270	55067	747	2.76
ボク	*	256087	792	24774	2396	3.02
おつ	*	485454	431	39277	1398	3.24
ノシ	*	352862	592	22786	2559	4.32

* indicates a Japanese slang, which is hardly to translate.

Table 2: Low R-ratio words (out-of-dictionary).



w		freq(w)	Rfreq(w)	user(w)	Ruser(w)	R-RATIO
ダウンロード	*download*	130200	1517	40373	1329	0.87
再起動	*reboot*	97634	1926	33090	1779	0.92
スイーツ	*sweets*	74842	2420	26448	2272	0.93
マック	*Mac*	231020	882	52983	821	0.93
ディズニー	*Disney*	42470	3072	17394	2920	0.95
プレゼン	*presentataion*	73507	2451	24264	2433	0.99
インストール	*install*	147180	1360	39767	1365	1
アカウント	*acount*	296177	690	55737	733	1.06
ＤＳ	*ＤＳ*	193599	1033	43858	1167	1.12

References

Burridge, K. and B. Kortmann (2008). Varieties of English: vol 3, Berlin and NY: Mouton de Gruyter.

Crowley, T. (2003). Standard English and the Politics of Language, Palgrave Macmillan.

Francis, W. N., H. Kučera and A. W. Mackie (1982). Frequency analysis of English usage: lexicon and grammar, Houghton Mifflin.

Freeborn, D. (2006). From Old English to Standard English: A Course Book in Language Variations Across Time, Palgrave Macmillan.

Ishino, H. (2000). "studies in the Japanese language " Kokugogaku **51**(3): 41-47.

Kurohashi, S., T. Nakamura, Y. Matsumoto and M. Nagao (1994). Improvements of Japanese Morphological Analyzer JUMAN. The International Workshop on Sharable Natural Language Resources.

NINJAL (1997). The total vocabulary and their written forms in ninety magazines of today.

Nishio, M., E. Iwabuchi and S. Mizutani (2009). IWANAMI Japanese dictionary 7th, Iwanamishoten.

Smith, J. (1996). An Historical Study of English: Function, Form and Change. London, Routledge.

West, M. (1953). A General Service List of English Words, Longman.