2013, Oct. 18. The 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)

Word in a Dictionary is used by Numerous Users

Is it true that the words in our dictionaries are really the "*standard*" language?

◎ Eiji Aramaki, Sachiko Maskawa, Mai Miyabe, Mizuki Morita, Sachi Yasuda ◎ Kyoto University/PRESTO eiji.aramaki@gmail.com

Dictionary editing requires enormous time of discussions to decide whether a word should be listed in a dictionary or not. So as to define a dictionary word, this study investigates that the number of word users for selecting dictionary words. In order to obtain the word users, we used about 0.25 billion tweets of 5-month period of approximately 100,000 people. This study compares the classification performance of various measures. The result of the experiment reveals that a word in a dictionary is used by numerous users.

Don't you think "*to google*" should be contained in a dictionary?

What about "*to search*"?

	"*to search*"	"*to google*"
(1) Word Frequency	low	high
(2) Usage Period	long	short
(3) **User Population**	small	big

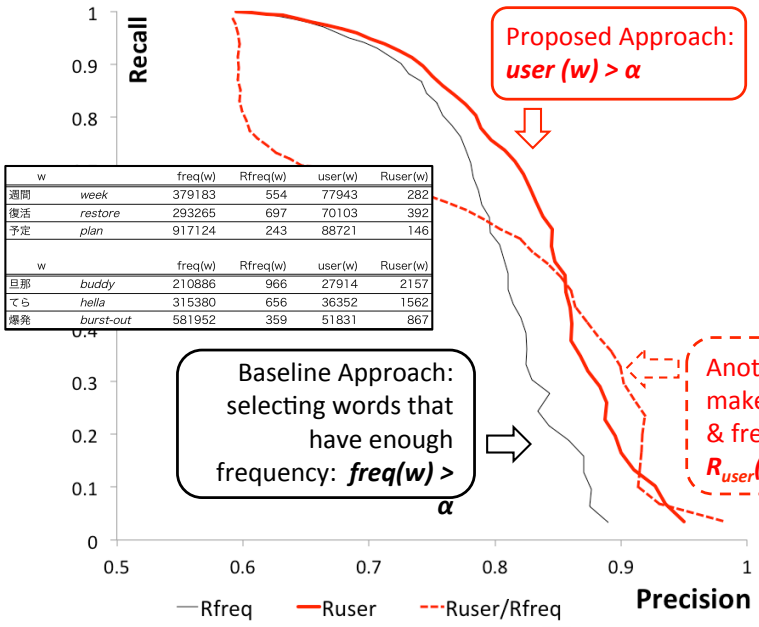
(3) **user population** has been hard to investigate. However, now we can see it from **SNS big-data**

Experiment:
input= Wikipedia entry names (4,000 nouns)
positive examples = 2,598 nouns (listed in dictionary*)
negative examples = 1,402 nouns (not listed in dictionary)

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

100,000 people tweets for 143 days
(from Nov. 2009 to Mar. 2010)
0.25 billion tweets (4,258,707,255 words)

freq(w): frequency of using word ***w***.
R_{freq}***(w)***: rank of ***freq (w)***.
user(w): number of users of a word ***w***.
R_{user}***(w)***: rank of ***user(w)***.

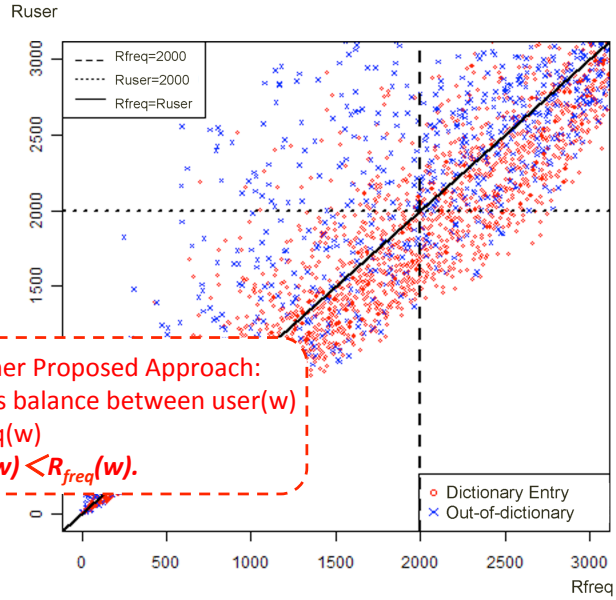


Proposed Approach: ***user (w) > α***

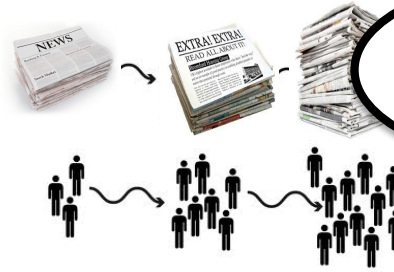
w		freq(w)	Rfreq(w)	user(w)	Ruser(w)
週間	*week*	379183	554	77943	282
復活	*restore*	293265	697	70103	392
予定	*plan*	917124	243	88721	146

w		freq(w)	Rfreq(w)	user(w)	Ruser(w)
旦那	*buddy*	210886	966	27914	2157
てら	*hella*	315380	656	36352	1562
爆発	*burst-out*	581952	359	51831	867

Baseline Approach: selecting words that have enough frequency: ***freq(w) > α***



Another Proposed Approach: makes balance between user(w) & freq(w)
$R_{user}(w) < R_{freq}(w)$.



Not the freq.!
BUT
The # of users is important!

w		freq(w)	Rfreq(w)	user(w)	Ruser(w)	Ruser/Rfreq
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	*MacDonald*	231020				0.93
ディズニー	*Disney*	42470				
プレゼン	*presentataion*	73507				
インストール	*install*	147				
アカウント	*acount*	29				
D S	*D S*	19				

This study tells us the word lists possibly in the future dictionaries!

Contact info: http://mednlp.jp