Synthesis of Secure Passwords

Tomoki Sato　Hiroaki Kikuchi

Graduate school of Engineering,
Tokai University

What is "*good*" password?

- A good password is composed of common words that are easy to type in
- A good password is an extraordinary phrase that is hardly ever used
  - These two requirements conflict with each other

"book and apple"   "Wingerdium Leviosa"

Existing study

- [1]Nishizaka et al."PIN authentication using Japanese password over cellular phone", IPSJ Tech Report, 2010.
  - Automatic password generation based on input method T9
  - Generated password is not always easy for humans to remember

Our Objective

- We propose a new synthesis method for good passwords that satisfy both requirements for good password

Our Approach

- Hypothesis

- If each of two words w_1 and w_2 has a high term frequency then the combination is not quite common

- Thus, the combined words gives strong impression

w_1	frequency
revolution	39 million

w_2	frequency
Granma	6.5 million

Combined word	frequency
revolution Granma	1

Our Contributions

1. New measure to evaluate degree of *impression*
2. New password synthesis scheme
3. Empirical study based on Google N-gram as a corpus

Formal Definitions

- Conflict *C*
  - *C* represents a degree how much reduction in frequency is given by combination of two words.
- Impression *I*
  - *I* is a measure based on subjective evaluation for words.
- Accuracy *A* (in remembrance)
  - *A* indicates how accurate subject can remember a given synthesized words for long term.

Conflict C

- Definition 2.1
  - A conflict of composition w_1w_2 is

$$C_x = -\frac{1}{10}\log\frac{S+1}{W_1+W_2}$$

- W_1, W_2, S : Frequency of w_1, w_2, Synthesized word
  - Frequency of word is defined in a set of web pages crawled by the search engine

$$C_x = -\frac{1}{10}\log\frac{1+1}{39,700,000+6,500,000} = 0.736$$

Example of conflict *C*

password	W_1	W_2	S	C
privacy festival	1.39×10^{7}	1.17×10^{7}	2	0.773
revolution Granma	3.97×10^{7}	6.5×10^{6}	1	0.736
eventually fill-in	1.69×10^{7}	3.74×10^{7}	6,630	0.391
first thought	1.5×10^{8}	1.69×10^{8}	54,300	0.377

Impression I

- Definition : impression for word x is

$$I_x = \frac{1}{n}\sum_{j=1}^{n} I_{x,j} - \bar{I}_j$$

- $I_{x,j}$ is a degree subjective impression of *j*-th test subject on word *x*
- $\bar{I}_j$ is average of all rating values evaluated by *j*-th subject
- The rating value range from 1(low) to 5(high)

Example of Impression /

password	subject1	subject2	subject3	Impression
privacy festival	5	5	3	2.05
revolution Granma	5	4	3	1.83
eventually fill-in	1	4	2	-0.28
first thought	1	2	2	-0.06
$\bar{I}_j$	3	3.75	2.5	

Accuracy *A*

- Definition 3.1
  - Accuracy of word x for short-term memory defined

$$A_x = \frac{1}{3n} \sum_j a_{j,x}$$

$$a_j = \begin{cases} 3 & \text{if first try is correct} \\ 2 & \text{if second try is correct} \\ 1 & \text{if third try is correct} \\ 0 & \text{if all tries are failures} \end{cases}$$

Example of Accuracy *A*

password	Subject1	Subject2	Subject3	Subject4	*A*
privacy festival	1	3	3	3	83.3%
revolution Granma	3	3	3	3	100%
eventually fill-in	0	3	3	0	50%
first thought	2	3	0	0	41.7%

Proposed Scheme

- Input : corpus

1. choose top 10,000 words in frequency from corpus(dataset),
2. classify the words into subsets, *noun*, *verb*, *adjective*, and *adverb*.
3. choose randomly two words from categories, (adverb + noun) or (noun + noun), and then grade pairs in conflict C.

- Output : synthesized passwords

Google N-gram

- A Japanese dataset extracted from web pages collected via a crawler
- It contains many words thats are very commonly used in the Internet

Example of Google N-gram

word	frequency
"capsule"	1,604,601
"horse"	2,967,320
"joint"	1,484,470

Experiment

1. Subjective evaluation of impression /
   - 18 subjects(students) evaluate passwords and answer impression degrees
2. Accuracy of remembrance
   - 16 subjects remember 4 synthesized passwords for each subject
   - 3 days later, they try a test to see how accurately they can remember 4 passwords

Fig. 1 : Impression / in terms of conflict *C*

correlation coefficient : 0.617

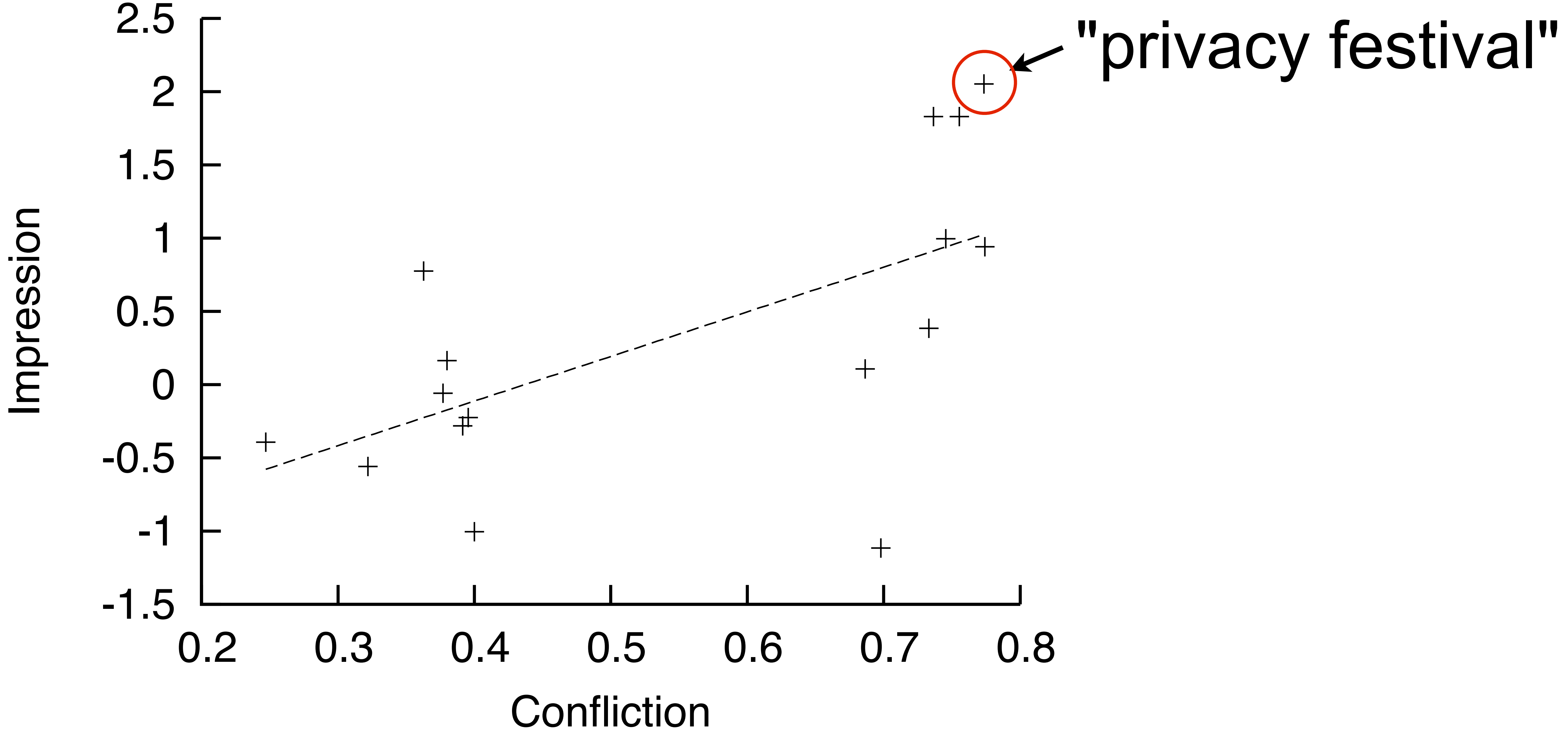


Fig. 2 : Accuracy *A* in terms of conflict *C*

correlation coefficient : 0.431

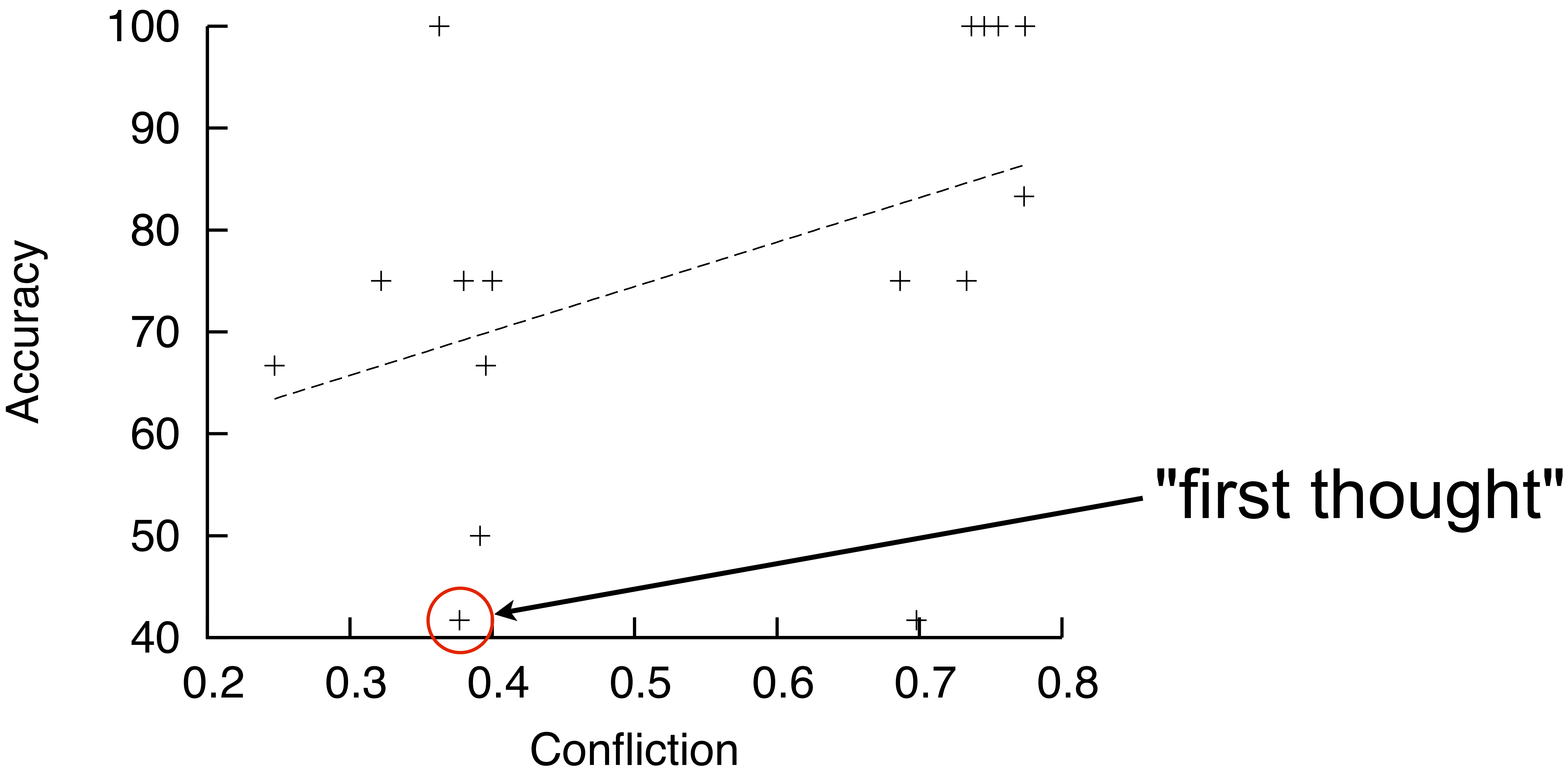
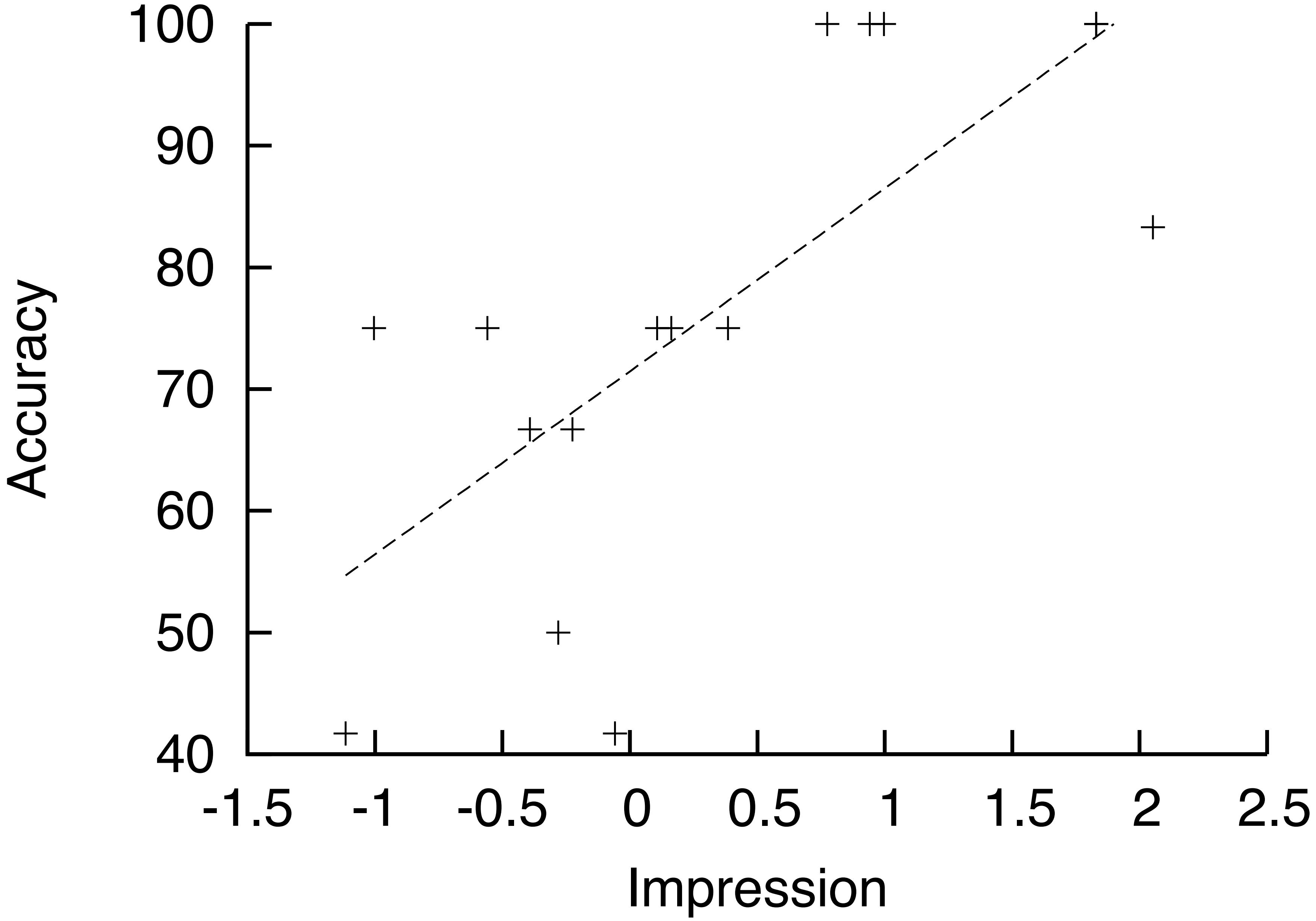


Fig. 3 : Accuracy *A* in terms of impression *I*

correlation coefficient : 0.733



Discussion

- In order to clarify the reason of failure

TRUE	answer	reason
privacy festival	private photo	similar words
eventually funny (可怪しい)	eventually susceptive (可怪しい)	homonym
first thought (初めて)	begin thought (はじめて)	Hiragana-Kanji conversion

Conclusion

- We have proposed a new way to synthesize good passwords that are easy to remember
- Our experiment shows a clear positive correlation between conflict C and impression I
- The synthesized passwords perform well in term of accuracy A in memory