Personalized Reading Support for Second-Language Web Documents by Collective Intelligence

IUI 2010, Feb. 2010, Hong Kong, China

The University of Tokyo

Yo Ehara,

Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa

http://en.newikipedia.org/

Second-language, reading support

Second-language (L2):

any language learned after the first language (cf. Wikipedia)

e.g. English for the speaker (me)

Among ways of *reading support*, we focus on:

- *Glossing* Web documents

  = to annotate words in a document with their meaning.

Gloss

Web document

Existing Glossing Systems

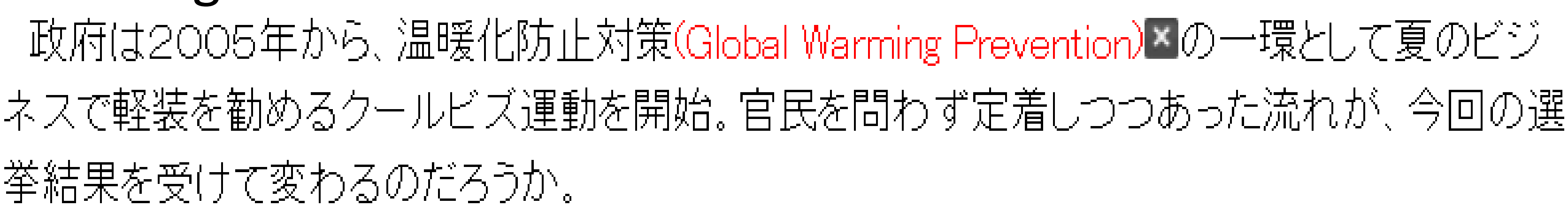
Show *glosses* (meaning) when words are clicked.

pop-jisyo (Coolest.com, 2001)

shows glosses in pop-ups

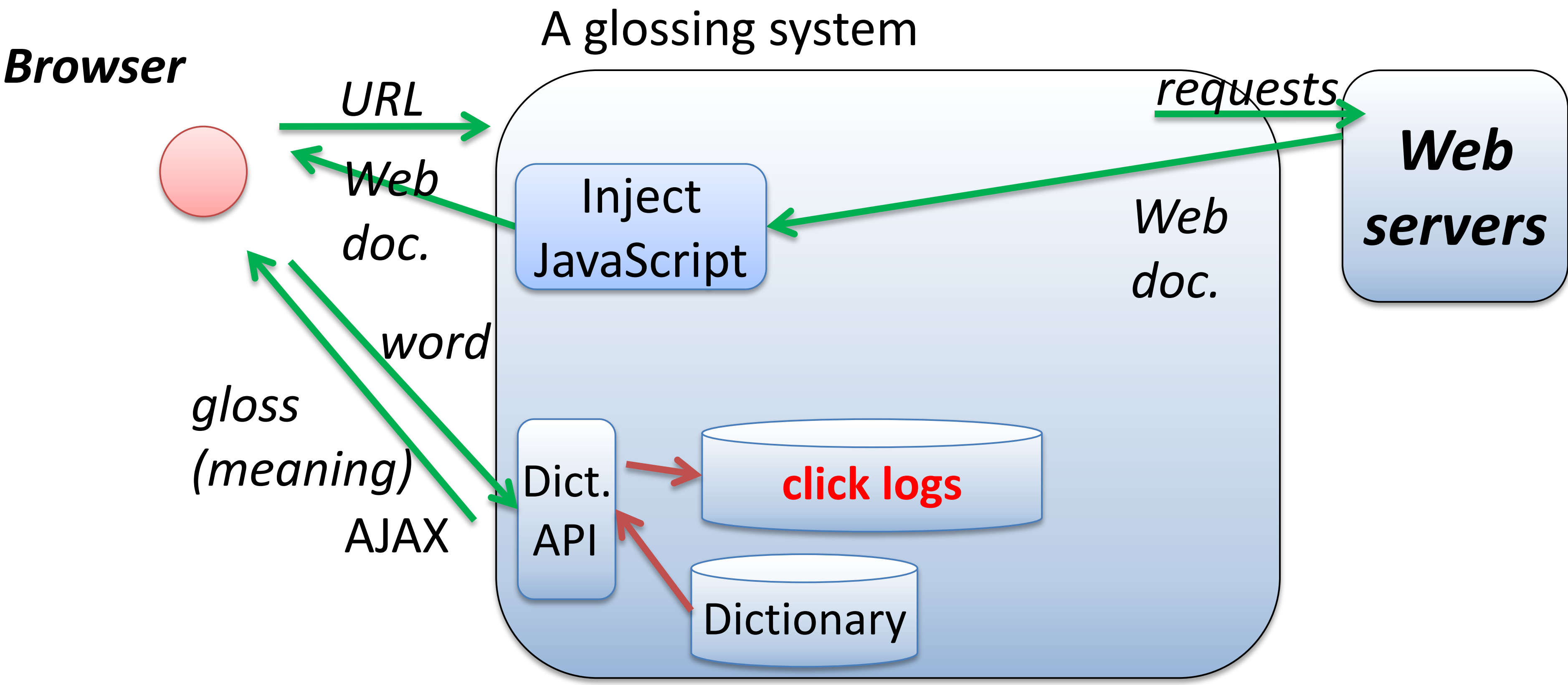
popIn (Cheng, 2008)

embeds glosses within a Web docuement



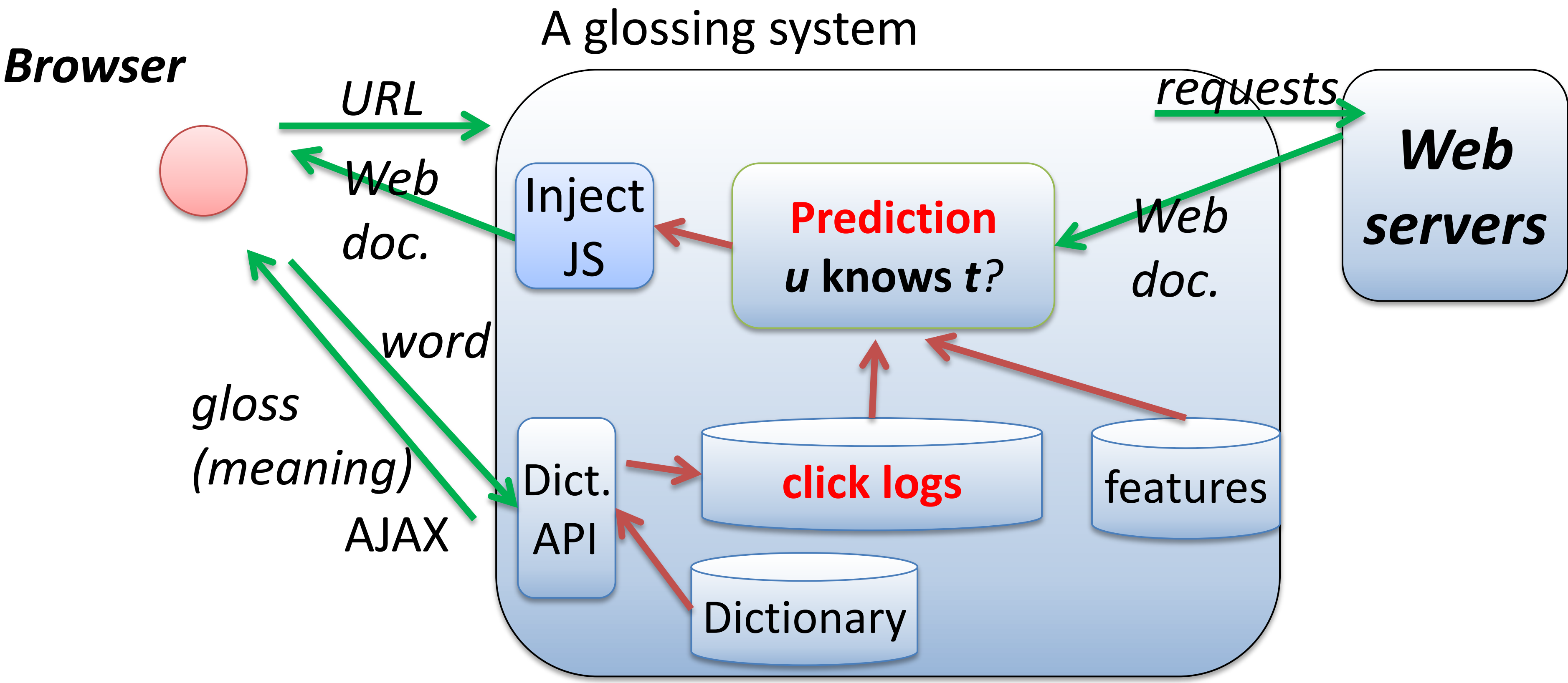
How they work?

By mediating between ***browsers*** and ***Web servers***



Though they tell which words the users don't know, click logs have been discarded

Proposal



Use click logs to predict ***"u* knows *t*"**? (*u*: user, *t*: word)

Users implicitly *collaborate* each other to train predictor

Demo

Access (currently supports only Japanese glosses) :

http://en.newikipedia.org/wiki/*item*

where *item* is the item of your interest in Wikipedia.

To avoid cumbersome log-in procedure:

- A browser is regarded as a user.
- Mean English ability is used for the first access.

Note that this system can support any Web pages.

Actually we once created a system that works on most of the English Web pages, http://www.socialdict.com/URL. In some Web pages, however, their JavaScript scripts didn't co-operate with our JavaScript script and resulted in corrupted display.

As this is not our main focus, we prepared a version limiting to Wikipedia and use this version as a demo.

Prediction : *personalized*

in the sense that prediction differs from user to user

yellow : Predicted to be *known to the user*

red : Predicted to be *unknown to the user*

Example from *the same* document:

low ability

When the crew asked if there were doctors (博士(の資格),医師,《口語》医師を開業する,治療をする,治療を受ける,薬を飲む,混ぜ物を入れる,(報告・証拠などを)ごまかす,不正な変更を加える,(機械などを)修理する,博士号,医者,内科医(physician),《口語》修理屋,修繕屋,(俗)(船内,キャンプの)料理人,先生) on the flight, (フライト,逃亡,飛行,階段,逃避,便) Dr Julien Struyven, 72, a cardiologist (心臓内科医) and radiologist ( ) from ((原料・材料)〜から,から) Brussels, went to the cockpit and examined (試験する,尋問する,調べる,検査する) the pilot. (パイロット,水先案内人,を導く,を通す,試験的な,操縦する,導く)

"He was not alive," (生存している,活動状態の,生きている,活発な) Dr Struyven told (見分ける,話す,言う,教える,しかる,数える,表われる,分かる) AP. There was "no chance (偶然の,偶然,幸運,可能性,危険,機会,偶然である,予期せぬ / Chances are that ~ : 恐らく〜であろう) at all" (すべての,すべての人[もの,こと],すべての人びと) of saving (1.節約する,救いの,節約(する),倹約(の),救済となる,倹約する,つましい,省力的な,2.救い,救援,救助,救済,省力,割引,値引率,3.〜以外は(except)) him, (彼を,彼に,彼) he said.

Dr Struyven said he suspected the pilot (パイロット,水先案内人,を導く,を通す,試験的な,操縦する,導く) had suffered (経験する,を被る,苦しむ,悩む,許す,受ける,損害を受ける) a cardiac arrest. (取り押さえる,逮捕する,つかまえる,引き止める,逮捕(する),止める) He said he used (利用,使う,利用する,(体,能力など)を働かす,使用,利用法) a defibrillator ( ) to try (1.〜を試験する,試みる,やってみる,試す,2.〜に苦難を与える,3.【法律】(事件を)審問する,審理する,裁判する,(人を)裁判にかける) and revive (生き返らせる,回復させる,元気にする,復興させる,蘇る,元気付く / The early dawn revived her.) the pilot, (パイロット,水先案内人,を導く,を通す,試験的な,操縦する,導く) but it was too (もまた,あまりに〜すぎて,あまりに) late.

Pilots are subject to rigorous medical checks (1.停止,妨害,検査,照合,小切手,反撃,勘定書,会計伝票,2.チェックする,急に止める,抑制する,調査する,一致する,3.(荷物などを)預ける) which increase in frequency with age. (年齢,時代)

high ability

When the crew asked if there were doctors on the flight, Dr Julien Struyven, 72, a cardiologist (心臓内科医) and radiologist ( ) from Brussels, went to the cockpit and examined (試験する,尋問する,調べる,検査する) the pilot.

"He was not alive," (生存している,活動状態の,生きている,活発な) Dr Struyven told AP. There was "no chance at all" (すべての,すべての人[もの,こと],すべての人びと) of saving him, (彼を,彼に,彼) he said.

Dr Struyven said he suspected the pilot had suffered a cardiac arrest. He said he used a defibrillator ( ) to try and revive (生き返らせる,回復させる,元気にする,復興させる,蘇る,元気付く / The early dawn revived her.) the pilot, but it was too late.

Pilots are subject to rigorous medical checks which increase in frequency with age.

Research Questions

1. To predict, *word difficulty* and *users' ability* needs to be estimated from the click logs.

   Can we estimate meaningfully? So that these measures are comparable to those used in *language testing*?

   – Yes. Language testing uses *IRT* model for these measures and we can use it for this task as well.

2. Can the system learn click logs dynamically (every time a user clicks)?

   – Yes. We can use SGD, an on-line algorithm, to train *IRT* model.

I will explain IRT and SGD

Item response theory (IRT)

Probabilistic models used in many *testing* studies including existing *language testing* like TOEFL.

Testing = estimate *difficulty* and *ability* from *test results*

= *click logs*

Rasch model : simplest version of IRT.

Notations:

User $u \in U$, Words $t \in T$, $y \in \{0,1\}$

y=1: u knows t, y=0: u doesn't know t

Accumulated click logs (y_n, u_n, t_n) :

$(y_1, u_1, t_1), (y_2, u_2, t_2), \ldots, (y_N, u_N, t_N)$

Rasch Model

Input：

$(y_1, u_1, t_1), (y_2, u_2, t_2), \ldots, (y_N, u_N, t_N)$

Parameters:

θ_u : u's ability

d_t : t's difficulty

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_u, \ldots, \theta_{|U|})^{\mathrm{T}}$$

$$\mathbf{d} = (-d_1, \ldots, -d_t, \ldots, -d_{|T|})^{\mathrm{T}}$$

Model：

$$P(y_n = 1 \mid u_n, t_n) = \sigma(\theta_u - d_t)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$ （sigmoid function）

Estimation: *ML* or MAP (prior on ability and difficulty)

$$\hat{\boldsymbol{\theta}}, \hat{\mathbf{d}} = \arg\max_{\boldsymbol{\theta}, \mathbf{d}} \prod_{n=1}^{N} P(y_n \mid u_n, t_n)$$

Rasch model to Log. Reg.

$$\mathbf{e}_u = (0,\ldots,1,0\ldots,0) \qquad \mathbf{e}_t = (0,\ldots,1,0\ldots,0)$$

1,... ,u,u+1,...,|U| *1,... ,t,t+1,...,|T|*

$$\mathbf{w}_{rasch} = \begin{pmatrix}\boldsymbol{\theta} & \mathbf{d}\end{pmatrix}^{\mathrm{T}}, \boldsymbol{\varphi}_{rasch}(u,t) = \begin{pmatrix}\mathbf{e}_u & \mathbf{e}_t\end{pmatrix}^{\mathrm{T}}$$

$$P(y_n = 1 | u_n, t_n) = \sigma\left(\theta_u - d_t\right)$$

$$= \sigma(\mathbf{w}_{rasch}{}^{\mathrm{T}} \boldsymbol{\varphi}_{rasch}(u,t))$$

inner product form

Logistic regression
a.k.a.:
log-linear model
Maximum entropy model

Answer to 1st Research Question

Can *word difficulty* & *user ability* be meaningful?
comparable to those used in *language testing*?

Here they are

Users' ability **Words' difficulty**

IRT:

$$\mathbf{w}_{rasch} = \left(\boxed{\boldsymbol{\theta}} \quad \boxed{\mathbf{d}} \right)^{\mathrm{T}}$$

$$\boldsymbol{\varphi}_{rasch}(u,t) = \left(\mathbf{e}_u \quad \mathbf{e}_t \right)^{\mathrm{T}}$$

LR (extended):

$$\mathbf{w}_{LR} = \left(\boxed{\boldsymbol{\theta}} \quad \boxed{\mathbf{d} \quad \mathbf{w}_a} \right)^{\mathrm{T}}$$

$$\boldsymbol{\varphi}_{LR}(u,t) = \left(\mathbf{e}_u \quad \mathbf{e}_t \quad \boxed{\boldsymbol{\varphi}_a} \right)^{\mathrm{T}}$$

Extra word feature

By adding extra word features, we can extend IRT with comparability remained

Extra features

Users' ability Words' difficulty

IRT:

$$\mathbf{w}_{rasch} = \left(\boxed{\boldsymbol{\theta}} \quad \boxed{\mathbf{d}}\right)^{\mathrm{T}}$$

$$\boldsymbol{\varphi}_{rasch}(u,t) = \left(\mathbf{e}_u \quad \mathbf{e}_t\right)^{\mathrm{T}}$$

LR (extended):

$$\mathbf{w}_{LR} = \left(\boxed{\boldsymbol{\theta}} \quad \boxed{\mathbf{d} \quad \mathbf{w}_a}\right)^{\mathrm{T}}$$

$$\boldsymbol{\varphi}_{LR}(u,t) = \left(\mathbf{e}_u \quad \mathbf{e}_t \quad \boxed{\boldsymbol{\varphi}_a}\right)^{\mathrm{T}}$$

Extra word feature

Extra word feature vector φ_a includes:

- **Google 1-gram**: word frequencies from a trillion Web documents
- **SVL12000**: manually annotated difficulty measure (1 – 12)

Training IRT to estimate parameters

Training consists of parameter *updates*.

- Batch learning [L-BFGS, Nocealdal+, 89], [Trust region Newton method, Lin+, 08]

  Converge to the *global optimum* as for Log. Reg.

  An update involves the *whole* training data.

- Online learning [**SGD** , Stochastic Gradient Descent]

  Not converge to the global optimum as for Log. Reg.

  An updates involves only the datum *that just has come*.

Answer to 2nd Research Question

Training consists of parameter *updates*.

- Batch learning [L-BFGS, Nocealdal+, 89], [Trust region Newton method, Lin+, 08]

  Converge to the *global optimum* as for Log. Reg.

  An update involves the *whole* training data.

- Online learning [**SGD** , Stochastic Gradient Descent]

  Not converge to the global optimum as for Log. Reg.

  An updates involves only the datum *that just has come*.

2nd RQ: Can the system learn click logs dynamically?

Answer: Yes if we use SGD.

Evaluation

Evaluation

Subjects: 16 university/graduate students

# of words answered: 12,000 *per a person*

1	**never seen the word before**
2	probably seen the word before
3	absolutely seen the word before but don’t know its meaning / tried to learn the word before but forgot its meaning
4	probably know the word’s meaning / able to guess the word’s meaning
5	absolutely know the word’s meaning

unknown

known

Evaluation Settings

Simulated the case a new user starts using our system from an accumulated log

\# of data in accumulated log: N_0

\# of data in the new user's log: N_1

- Data set:

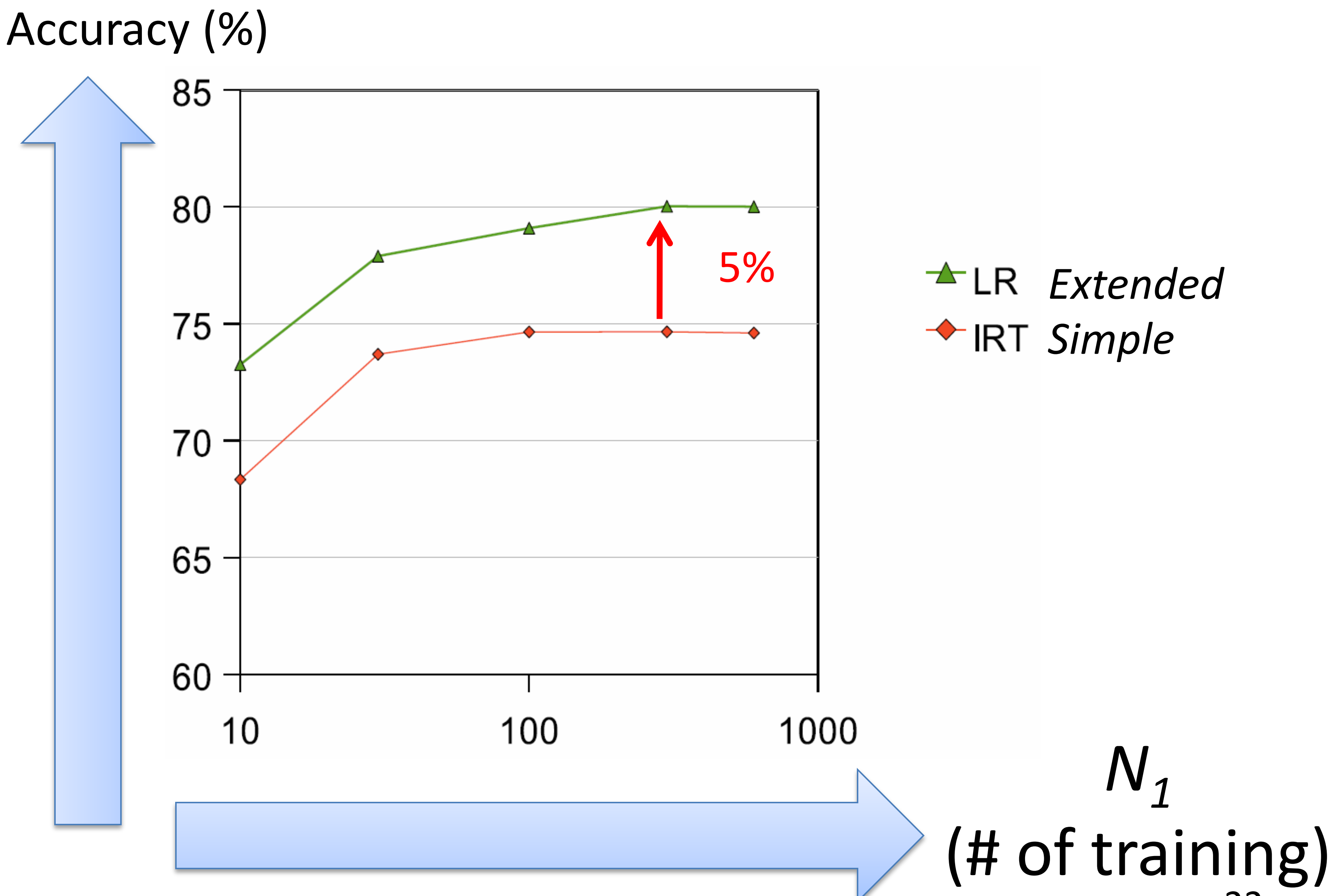
N_0+N_1 words (10 <=N_1 <=600)	*Training*
1400 words	*Development*
9999 words	*Test*

Accuracy = Ratio of words correctly predicted in *Test.*
Averaged over the 16 subjects

- smart.fm log is used as the accumulated log.

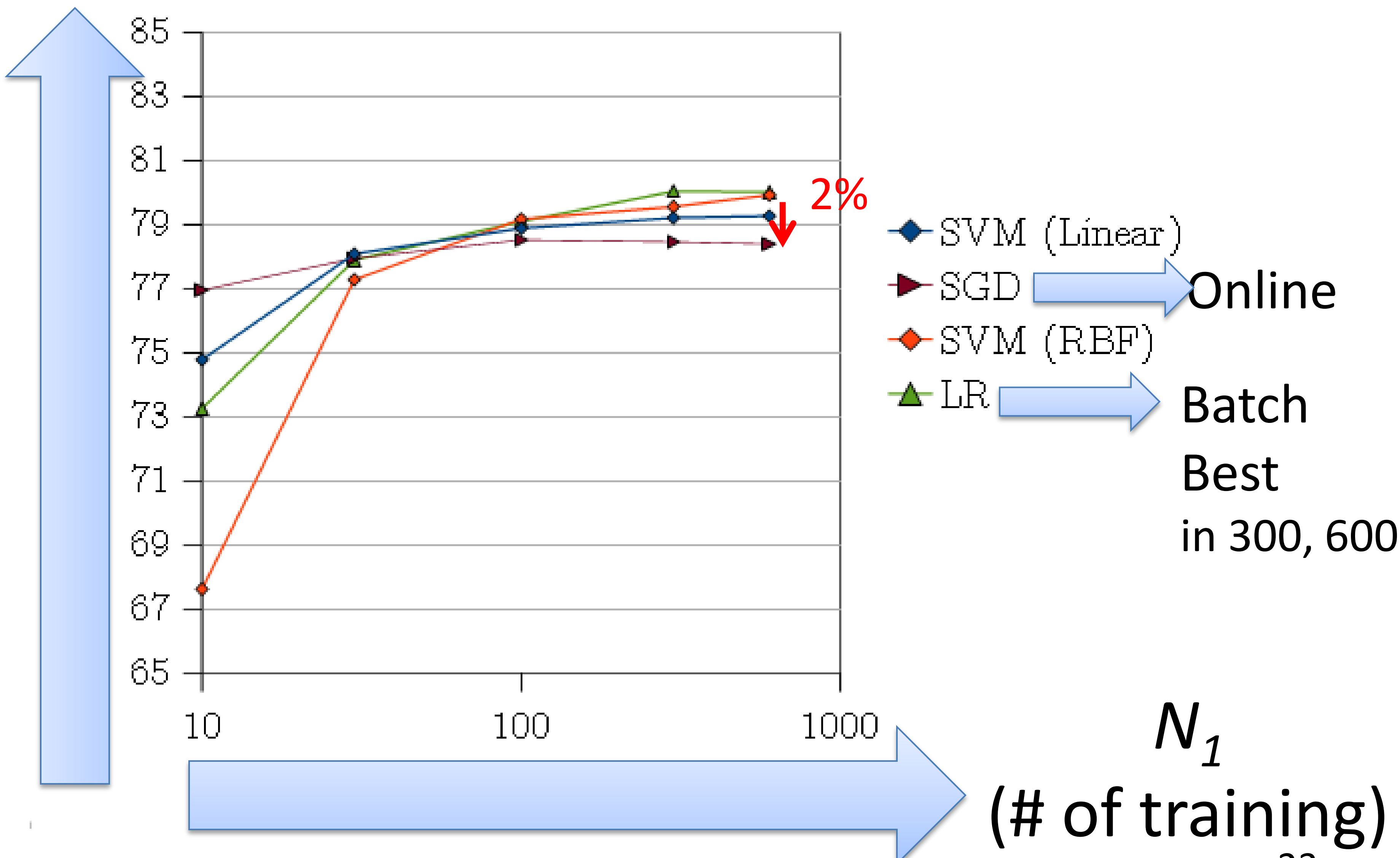
smart.fm is a system whose log stores millions of (y_n, u_n, t_n).

Effect by adding extra features



Effect by use of online learning

Accuracy (%)



N_1

(# of training)

Conclusions (Contributions)

- Invented a glossing system with *personalized* prediction that tells *who* knows *which* word

  by utilizing click logs having been discarded so far.
- *1st RQ*: Among binary classifications (e.g. SVM),

  use of *IRT* (Log. reg.) is preferable for this task since its measures (ability & difficulty) are comparable to those used in *language testing*.
- Extended IRT by adding extra word features and marked about 5% higher accuracy
- *2nd RQ*: SGD enables on-line learning of IRT and learns click logs dynamically with sacrifice of 2% acc.

Thank you for listening!

Aftermath of this presentation

1.: This work is accepted by *ACM Transactions on Intelligent Systems and Technology, Special issue on Tutoring and Coaching System*.

In the journal version, I simulated the case the 16 users read the 500 docs in the Brown corpus and showed, by using this system, that:

- the users can read more documents
  - (existing researches showed that, to read a document satisfactorily, a reader should know its 95% of words in occurrence.)
- the number of the users' clicks decrease

compared to the case the users simply click and look up every unfamiliar word.

2.: I collected the logs 3 times larger than mentioned in this study from smart.fm, which is closed to the public now. Modeling and analyzing this logs will be my further research.