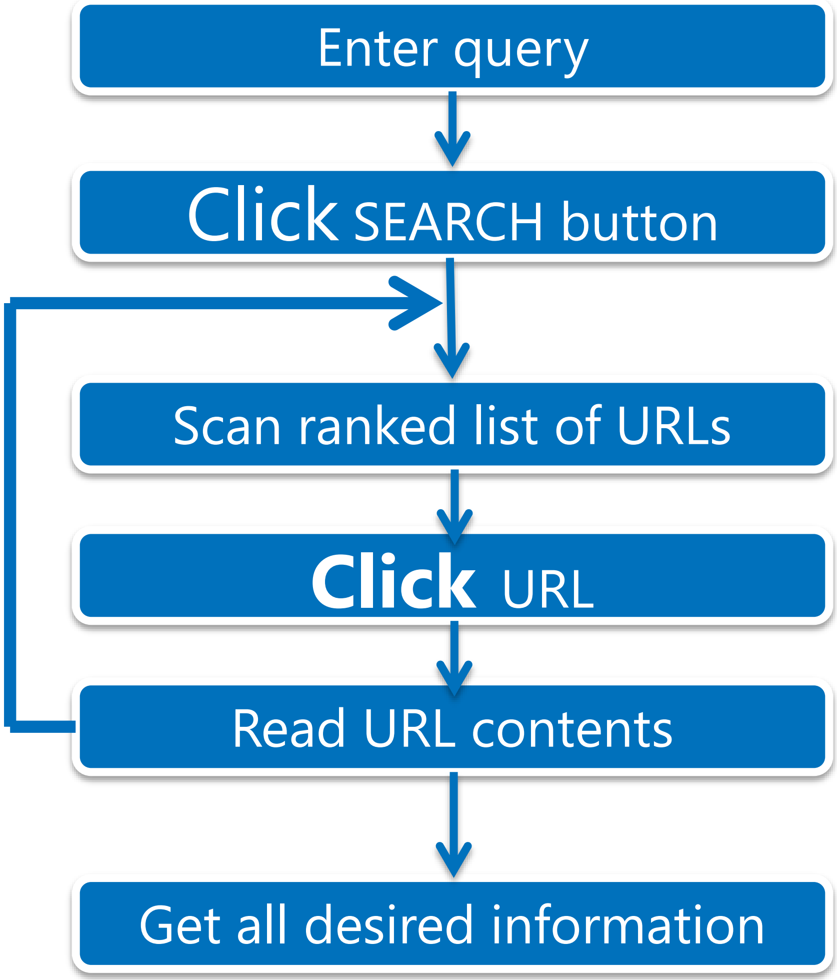
Two-layered Summaries for Mobile Search:
Does the Evaluation Measure Reflect User Preferences?

Makoto P. Kato (Kyoto U.), Tetsuya Sakai (Waseda U.),
Takehiro Yamamoto (Kyoto U.), Virgil Pavlu (Northeastern U.),
and Hajime Morita (Kyoto U.)

MOTIVATION AND TASK

IR Systems in *Ten-Blue-Link* Paradigm



Enter query

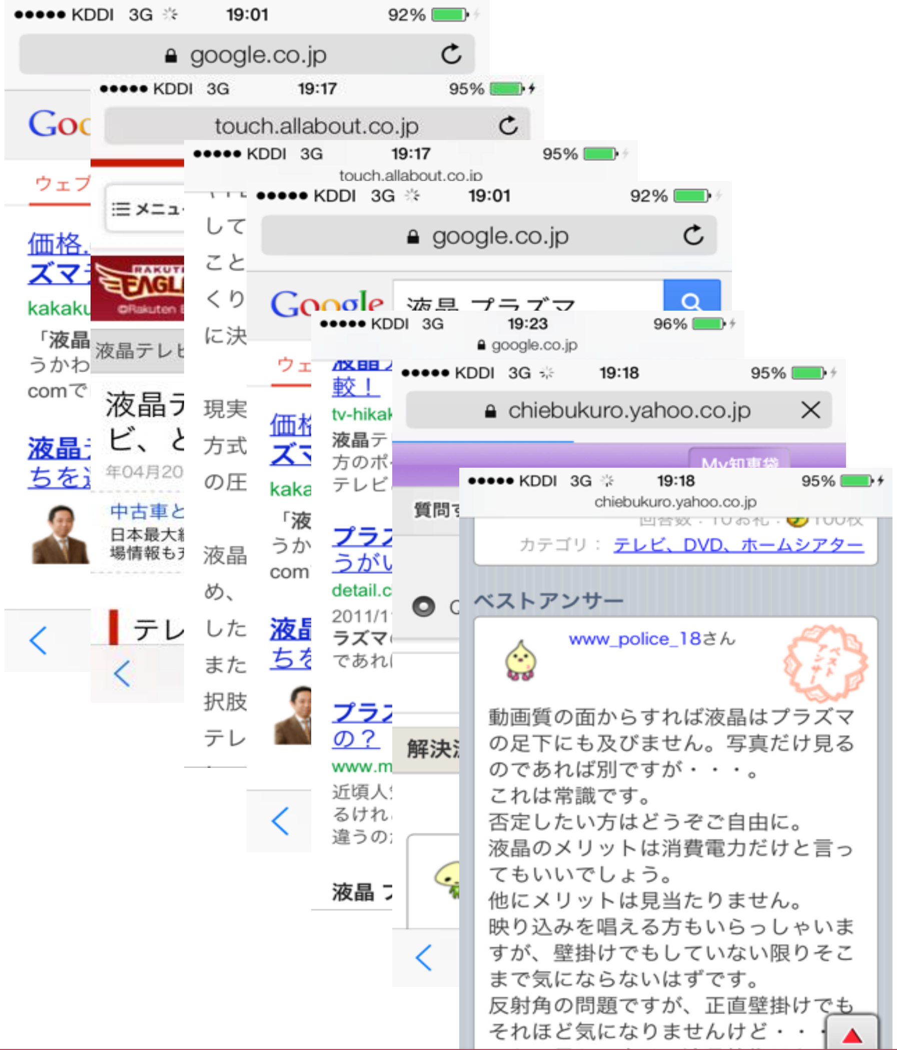
Click SEARCH button

Scan ranked list of URLs

Click URL

Read URL contents

Get all desired information



google.co.jp

touch.allabout.co.jp

google.co.jp

chiebukuro.yahoo.co.jp

●●●●● KDDI 3G 19:18 95%

chiebukuro.yahoo.co.jp

カテゴリ： テレビ、DVD、ホームシアター

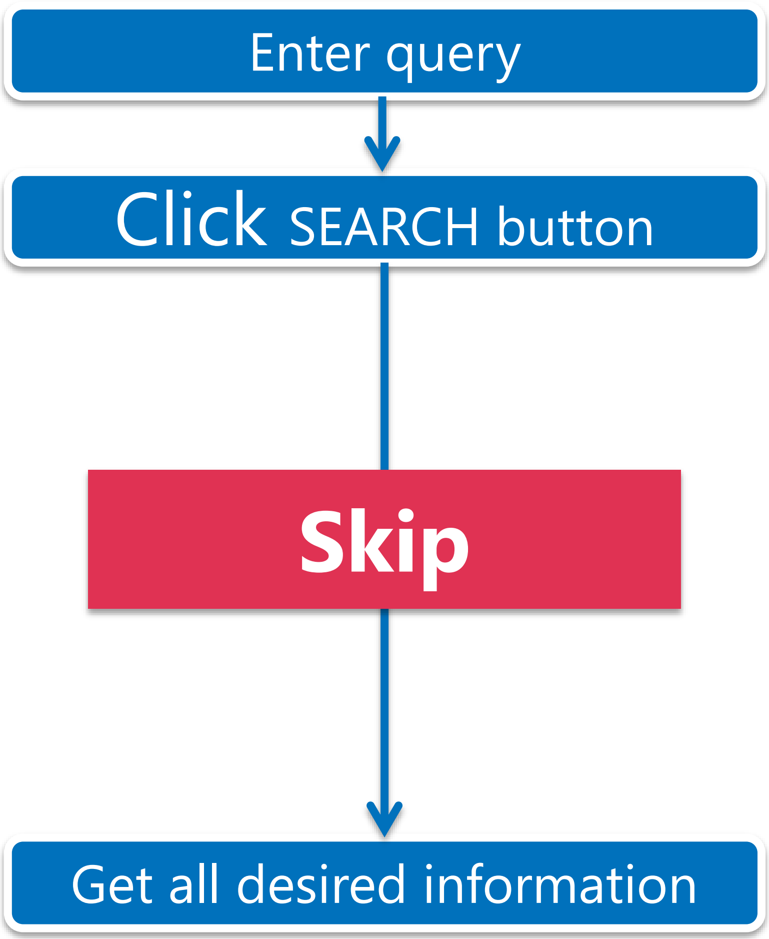
ベストアンサー

www_police_18さん

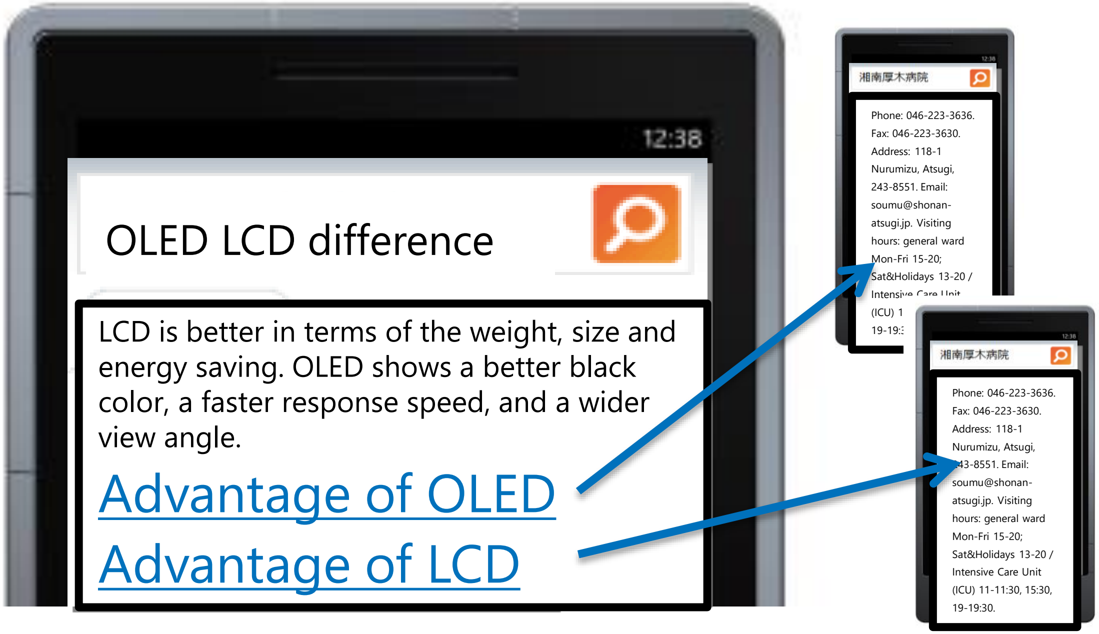
動画質の面からすれば液晶はプラズマの足下にも及びません。写真だけ見るのであれば別ですが・・・。
これは常識です。
否定したい方はどうぞご自由に。
液晶のメリットは消費電力だけと言ってもいいでしょう。
他にメリットは見当たりません。
映り込みを唱える方もいらっしゃいますが、壁掛けでもしていない限りそこまで気にならないはずです。
反射角の問題ですが、正直壁掛けでもそれほど気になりませんけど・・・

Long way to get all desired information

MobileClick System



System output


12:38

OLED LCD difference

LCD is better in terms of the weight, size and energy saving. OLED shows a better black color, a faster response speed, and a wider view angle.

Advantage of OLED

Advantage of LCD

12:38

湘南厚木病院

Phone: 046-223-3636. Fax: 046-223-3630. Address: 118-1 Nurumizu, Atsugi, 243-8551. Email: soumu@shonan-atsugi.jp. Visiting hours: general ward Mon-Fri 15-20; Sat&Holidays 13-20 / Intensive Care Unit (ICU) 1
19-19:3

12:38

湘南厚木病院

Phone: 046-223-3636. Fax: 046-223-3630. Address: 118-1 Nurumizu, Atsugi, 243-8551. Email: soumu@shonan-atsugi.jp. Visiting hours: general ward Mon-Fri 15-20; Sat&Holidays 13-20 / Intensive Care Unit (ICU) 11-11:30, 15:30, 19-19:30.

Task: Given a search query, return a two-layered textual output

Go beyond the "ten-blue-link" paradigm, and tackle *information* retrieval rather than document retrieval

iUnit Summarization Subtask at NTCIR-12

- **Given a query, a set of iUnits, and a set of intents, generate a two-layered summary**

Input: **Query**

NTCIR

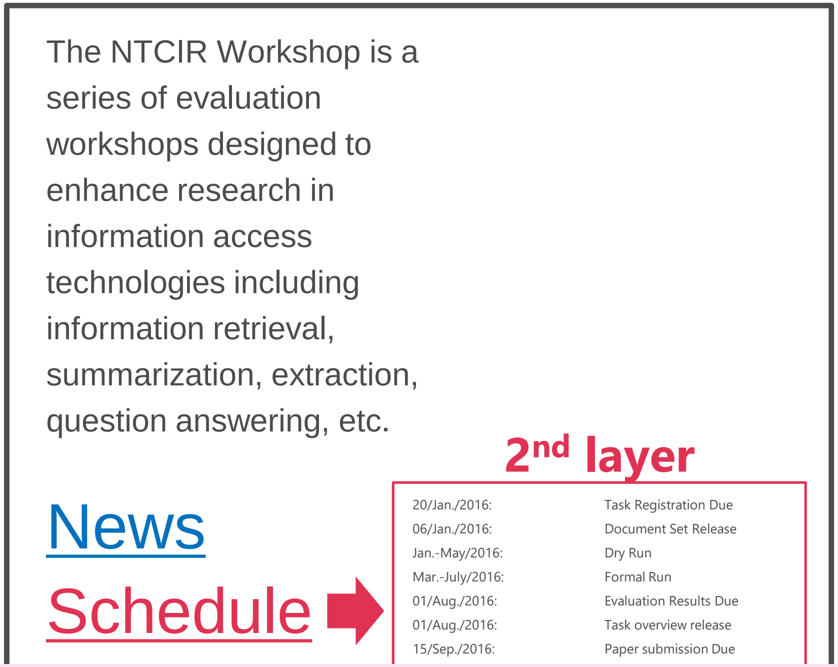
Input: **iUnit set**

iUnit
A series of evaluation workshops
Designed to enhance IA research
…

Input: **Intents**

Intents

Output: **Two-layered summary**



The NTCIR Workshop is a series of evaluation workshops designed to enhance research in information access technologies including information retrieval, summarization, extraction, question answering, etc.

News

Schedule

2nd layer

20/Jan./2016:	Task Registration Due
06/Jan./2016:	Document Set Release
Jan.-May/2016:	Dry Run
Mar.-July/2016:	Formal Run
01/Aug./2016:	Evaluation Results Due
01/Aug./2016:	Task overview release
15/Sep./2016:	Paper submission Due

Evaluation metric designed for mobile information access

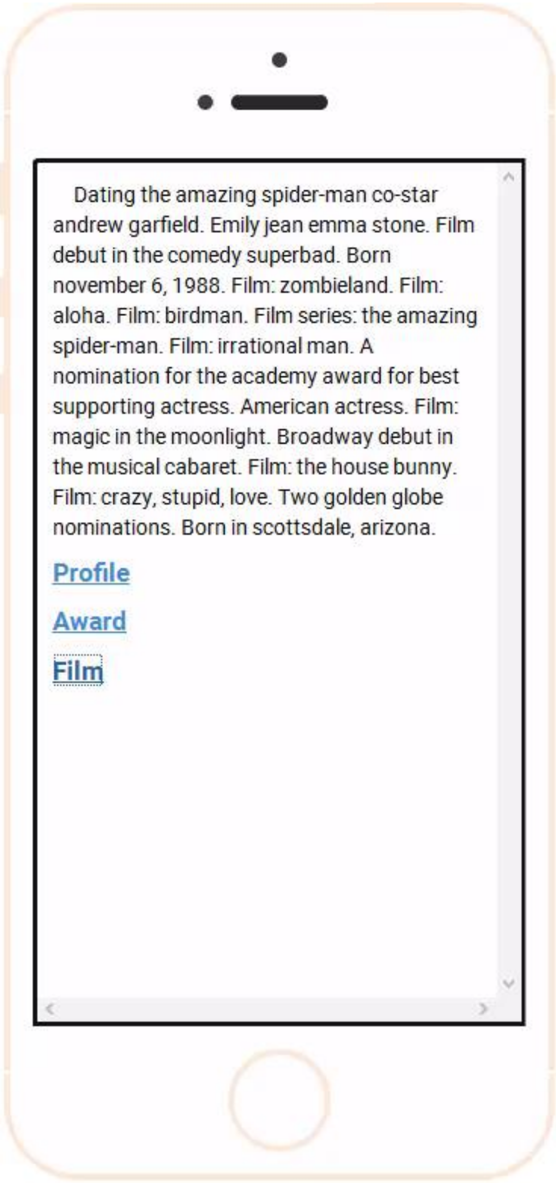
M-measure

0.5

Challenge

Lay out iUnits so that any types of users can be immediately satisfied

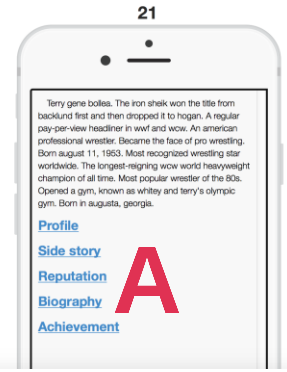
Two-layered Summary in Action



Dating the amazing spider-man co-star andrew garfield. Emily jean emma stone. Film debut in the comedy superbad. Born november 6, 1988. Film: zombieland. Film: aloha. Film: birdman. Film series: the amazing spider-man. Film: irrational man. A nomination for the academy award for best supporting actress. American actress. Film: magic in the moonlight. Broadway debut in the musical cabaret. Film: the house bunny. Film: crazy, stupid, love. Two golden globe nominations. Born in scottsdale, arizona.

Research Question Addressed in This Work

Does the Evaluation Measure Reflect User Preferences?



Terry gene bollea. The iron sheik won the title from backlund first and then dropped it to hogan. A regular pay-per-view headliner in wwf and wcw. An american professional wrestler. Became the face of pro wrestling. Born august 11, 1953. Most recognized wrestling star worldwide. The longest-reigning wcw world heavyweight champion of all time. Most popular wrestler of the 80s. Opened a gym, known as whitey and terry's olympic gym. Born in augusta, georgia.

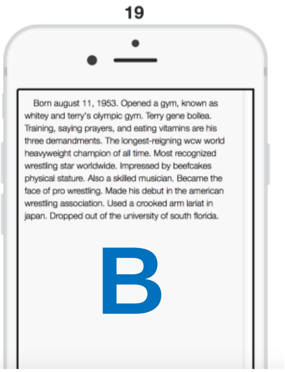
Profile

Side story

Reputation

Biography

Achievement



Born august 11, 1953. Opened a gym, known as whitey and terry's olympic gym. Terry gene bollea. Training, saying prayers, and eating vitamins are his three demandments. The longest-reigning wcw world heavyweight champion of all time. Most recognized wrestling star worldwide. Impressed by beefcakes physical stature. Also a skilled musician. Became the face of pro wrestling. Made his debut in the american wrestling association. Used a crooked arm lariat in japan. Dropped out of the university of south florida.

M-measure

0.5	0.4

Which is higher?

0.5 > 0.4

Which is better?

A > B

User preference

(# of users who prefer to A (B))

10	4

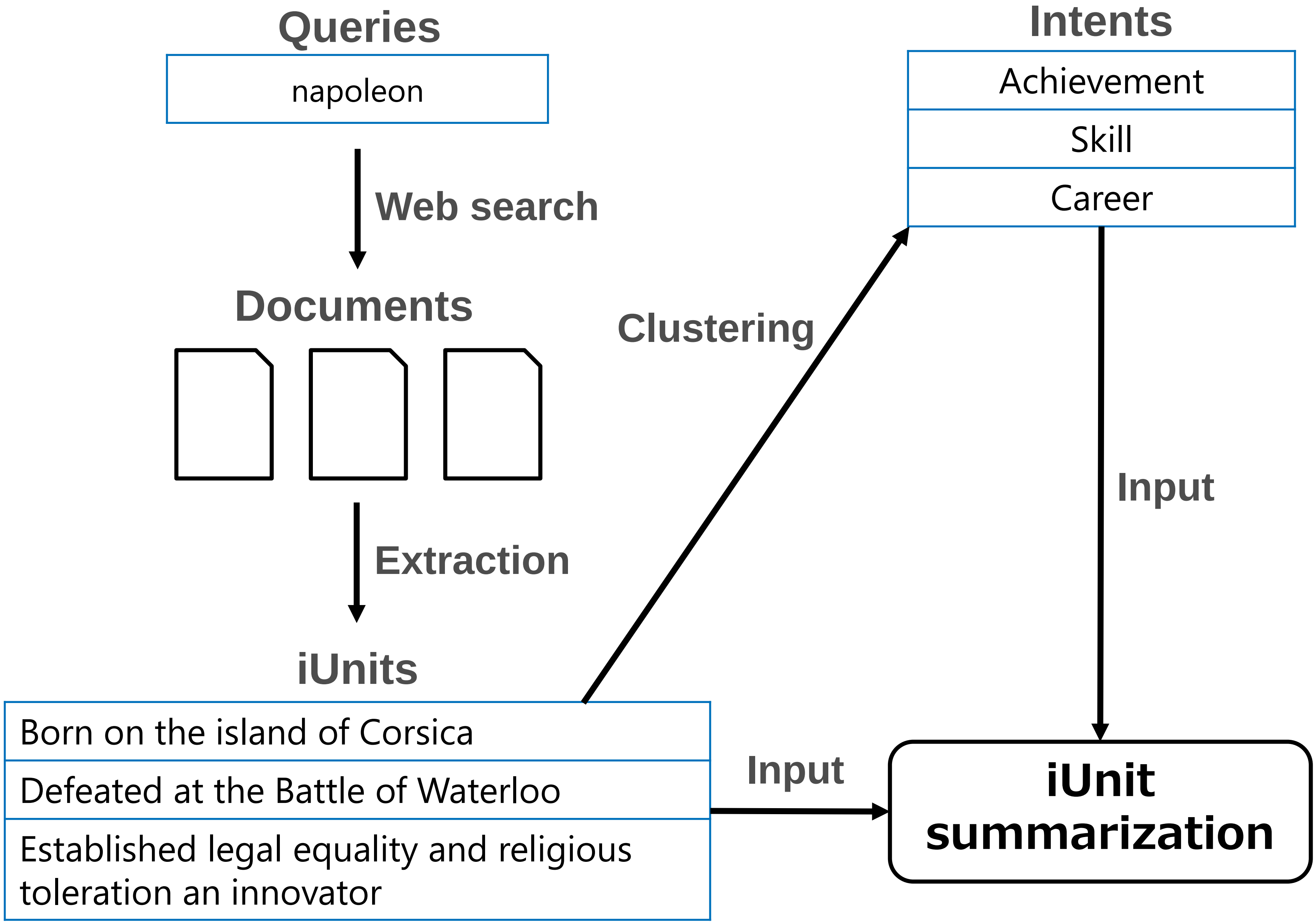
10 > 4

A > B

|| **Same?**

DATA

Overview of Data



Queries

napoleon

Web search

Documents

Extraction

iUnits

Born on the island of Corsica

Defeated at the Battle of Waterloo

Established legal equality and religious toleration an innovator

Clustering

Intents

Achievement

Skill

Career

Input

Input

iUnit summarization

Queries and Documents

- **Queries**
  - 100 English/Japanese queries
  - Most of which were ambiguous/underspecified
  - **Selected from five categories:**
    celebrity, location, definition, and QA (similar to NTCIR 1CLICK-2)

Examples

CELEBRITY	LOCATION	DEFINITION	QA
hulk hogan	bank adelanto	bitcoin	what is mirror made of
bruno mars	cafe killeen	divers disease	how to cook coleslaw
sharon stone	cincinnati art museum	windows 7	role of animal tail

- **Documents**
  - 500 commercial search engine results for each query
    from which iUnits were extracted

iUnits

- **Definition**
  - – Atomic information pieces relevant to a given query
- The number of iUnits
  - – **2,317** (23.8 iUnits per query) for English
  - – **4,169** (41.7 iUnits per query) for Japanese

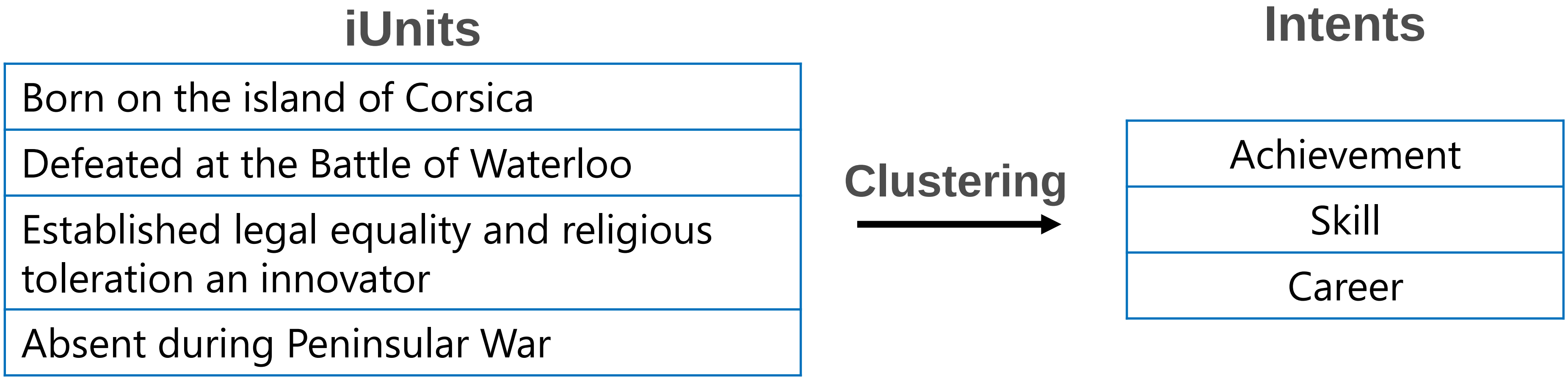
Examples of iUnits for query “Napoleon”

Born on the island of Corsica	General of the Army of Italy
Defeated at the Battle of Waterloo	One of the most controversial political figures won at the Battle of Wagram
Established legal equality and religious toleration an innovator	Baptised as a Catholic
Absent during Peninsular War	Cut off European trade with Britain

Intents

- **An intent can be defined as**
  - **A specific interpretation of an ambiguous query** (“Mac OS” and “car brand” for “jaguar”), or
  - **An aspect of a faceted query** (“windows 8” and “windows 10” for “windows”)

- **Obtained by clustering iUnits**



iUnits
Born on the island of Corsica
Defeated at the Battle of Waterloo
Established legal equality and religious toleration an innovator
Absent during Peninsular War

Clustering

Intents
Achievement
Skill
Career

EVALUATION

Per-intent iUnit Importance and Intent Probability

- **Importance of iUnits in terms of an intent**

In terms of intent "Definition"

iUnit	Importance
A series of evaluation workshops	**5**
Task Registration Due 20/Jun./2016	**3**

In terms of intent "Schedule"

iUnit	Importance
A series of evaluation workshops	**2**
Task Registration Due 20/Jun./2016	**5**

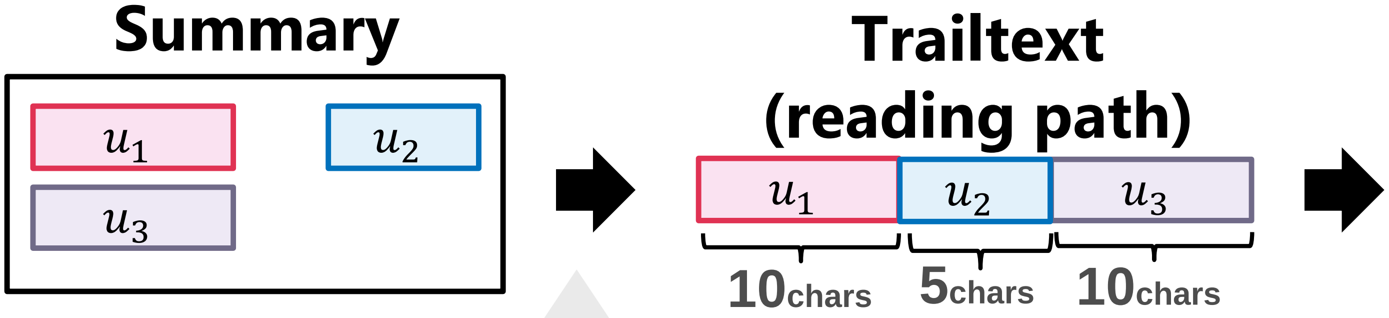
- **Intent probability** P(i|q)
  - Probability of having intent i for a given query q

Intent	Prob.
Definition	**0.4**
Schedule	**0.3**
Tasks	**0.3**

For details, see our MobileClick-2 overview paper

Evaluation of iUnit Summarization (Single-layer Case)

- Consider single-layered summary evaluation
- U-measure [Sakai and Dou. SIGIR2013]
  - Higher if more important iUnits appear earlier



U-measure

$G(u_1)(1-10/L)$
$+ G(u_2)(1-15/L)$
$+ G(u_3)(1-25/L)$

Create a list of iUnits by assuming that users read text from left to right, from top to bottom

$$U = \sum_{r=1} G(u_r)\left(1 - \frac{\mathrm{pos}(u_r)}{L}\right)$$

u_r: r-th iUnit
$G(u)$: importance of u
$\mathrm{pos}(u)$: offset of u from the beginning
L: patience parameter

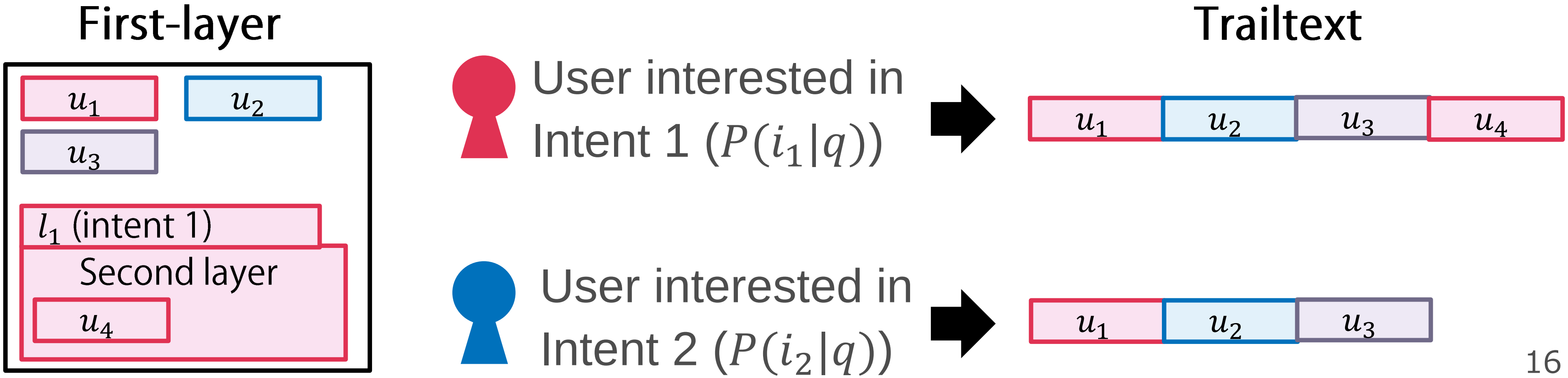
M-measure

- **M-measure**
  - **Expectation of U-measure over multiple *trailtexts***

$$M = \sum_{\mathbf{t}} P(\mathbf{t})U(\mathbf{t})$$

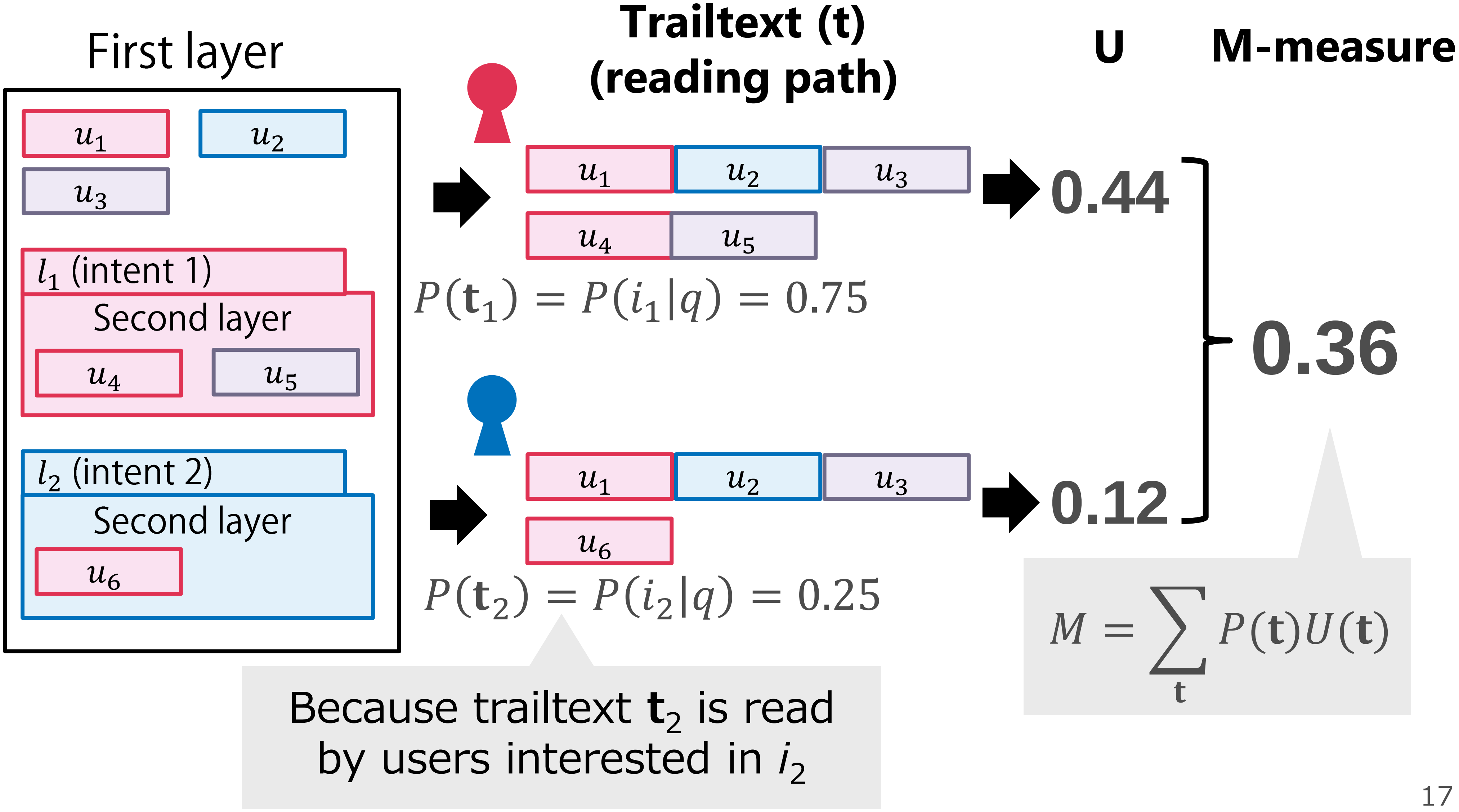
$P(\mathbf{t})$: probability of trailtext **t**
$U(\mathbf{t})$: U-measure of trailtext **t**

1. **Generate trailtexts by assuming that**
   - Users read a summary from the top of the first layer
   - Users click on an intent if they are interested in it



Evaluation of iUnit Summarization (Two-layer Case)

2. Compute the expectation of U-measure



First layer

u_1 u_2 u_3

l_1 (intent 1)

Second layer

u_4 u_5

l_2 (intent 2)

Second layer

u_6

Trailtext (t)
(reading path)

U

M-measure

u_1 u_2 u_3 u_4 u_5 → 0.44

$$P(\mathbf{t}_1) = P(i_1|q) = 0.75$$

u_1 u_2 u_3 u_6 → 0.12

$$P(\mathbf{t}_2) = P(i_2|q) = 0.25$$

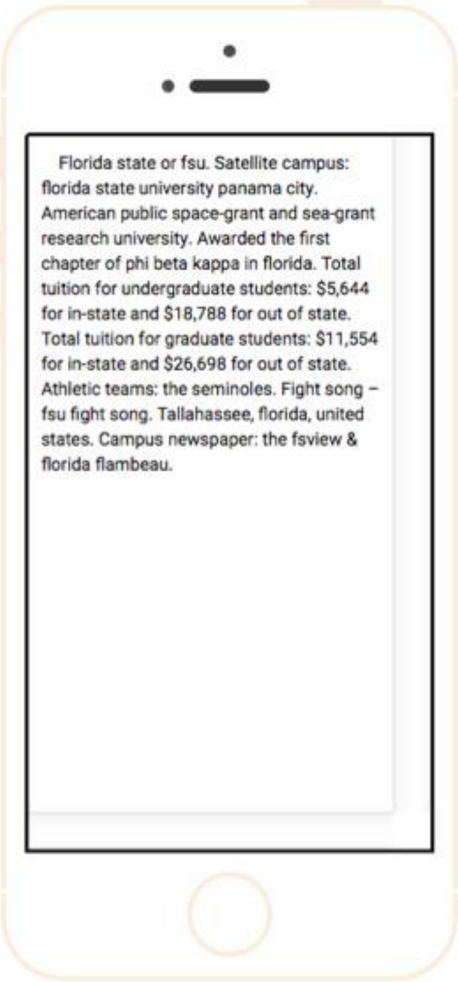
0.36

Because trailtext $\mathbf{t}_2$ is read by users interested in i_2

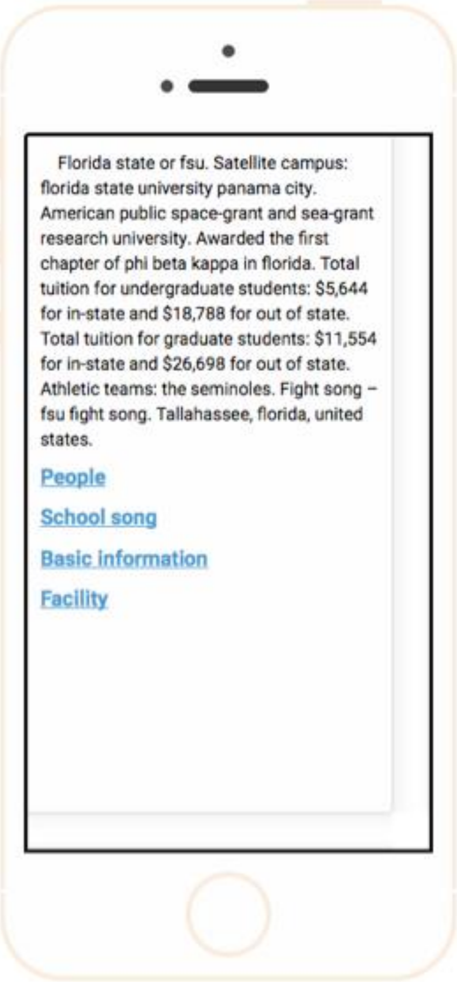
$$M = \sum_{\mathbf{t}} P(\mathbf{t})U(\mathbf{t})$$

EXPERIMENT

Pairwise Comparison



Florida state or fsu. Satellite campus: florida state university panama city. American public space-grant and sea-grant research university. Awarded the first chapter of phi beta kappa in florida. Total tuition for undergraduate students: $5,644 for in-state and $18,788 for out of state. Total tuition for graduate students: $11,554 for in-state and $26,698 for out of state. Athletic teams: the seminoles. Fight song – fsu fight song. Tallahassee, florida, united states. Campus newspaper: the fsview & florida flambeau.



Florida state or fsu. Satellite campus: florida state university panama city. American public space-grant and sea-grant research university. Awarded the first chapter of phi beta kappa in florida. Total tuition for undergraduate students: $5,644 for in-state and $18,788 for out of state. Total tuition for graduate students: $11,554 for in-state and $26,698 for out of state. Athletic teams: the seminoles. Fight song – fsu fight song. Tallahassee, florida, united states.

People

School song

Basic information

Facility

All possible pairs of 7 summaries for 25 queries were presented to about 14 users

Instruction in Pairwise Comparison

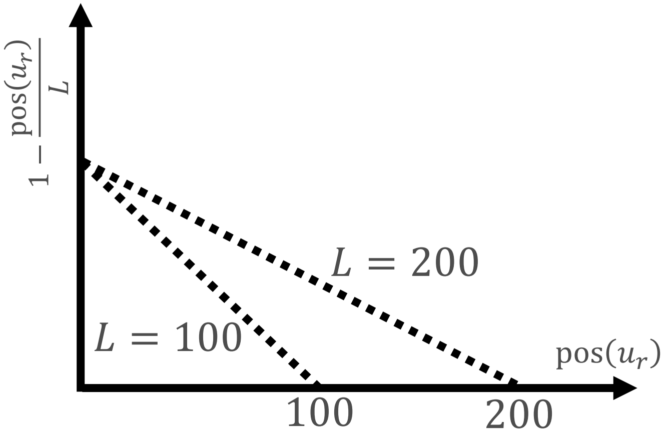
- **Users were asked to select either**
  the left one is better*, *the right one is better*, *equally good*, or *equally bad

- **Criteria:**

  (1) How much useful information you can get from the summary, and
  (2) How quickly you can get useful information from the summary

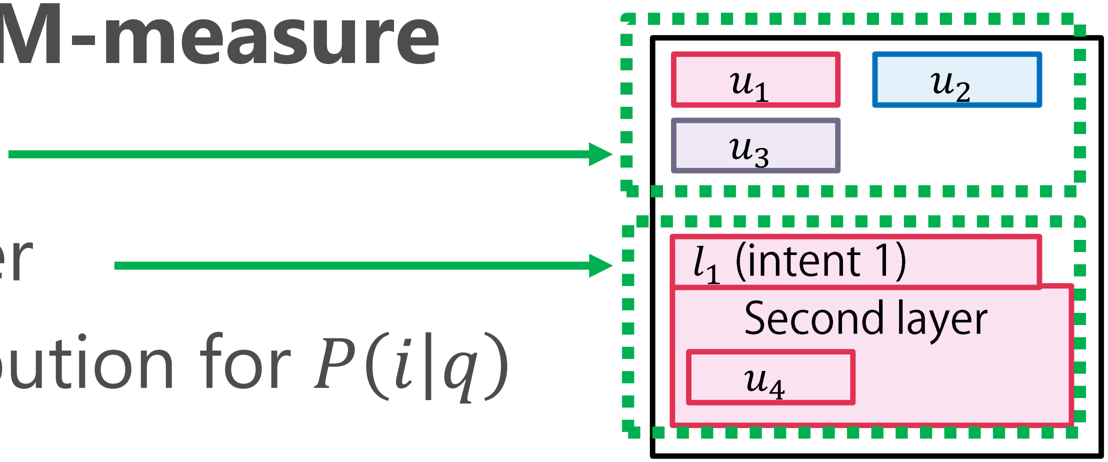
Settings of M-measure

- L **of U-measure in M-measure**
  - $U = \sum_{r=1} G(u_r) \max\left(0, 1 - \frac{\mathrm{pos}(u_r)}{L}\right)$



  - L is a patience parameter that controls how the gain of iUnits decreases as the user reads the text

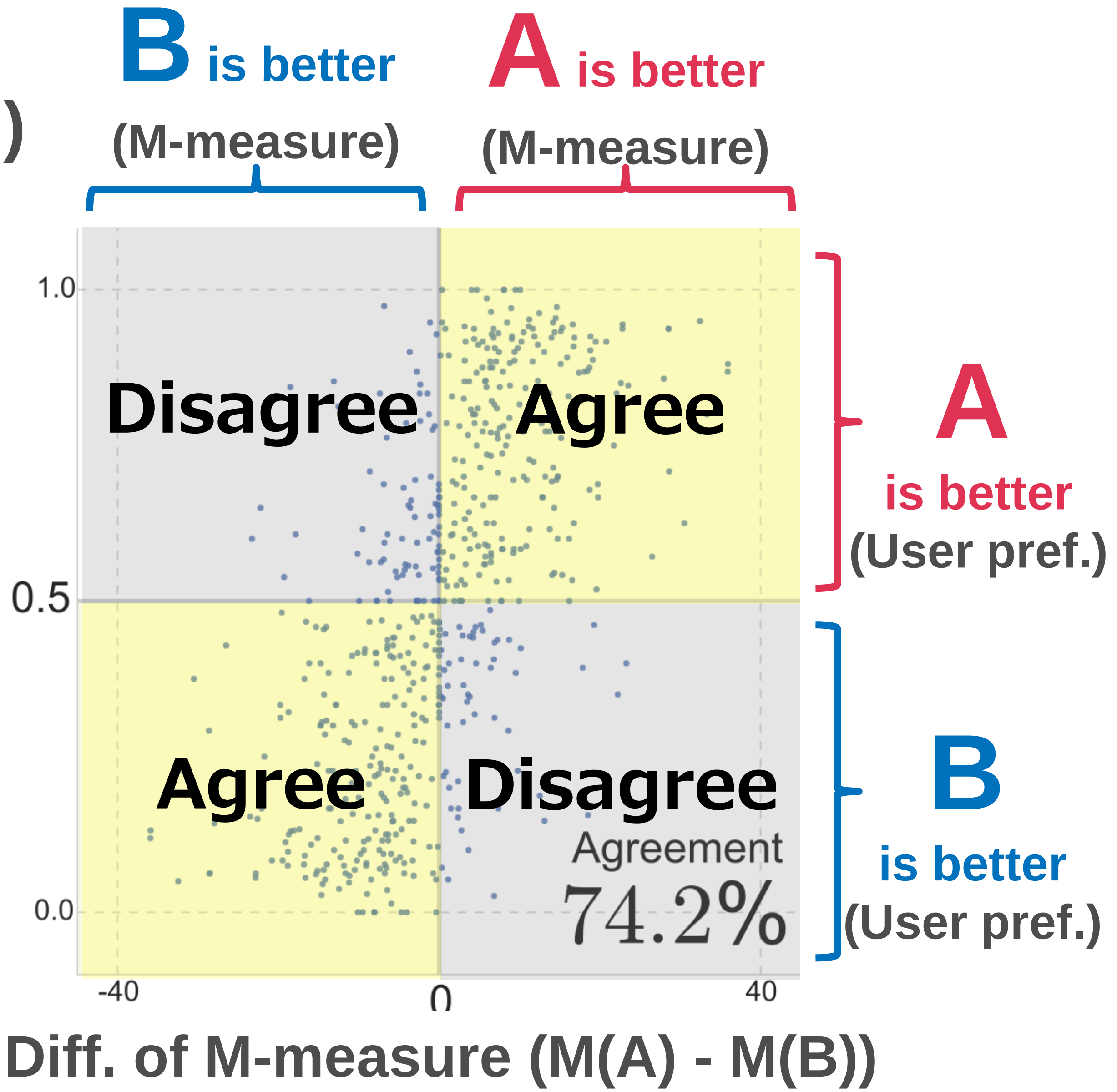
- **Simple variants of M-measure**
  - Use only first layer
  - Use only second layer
  - Use a uniform distribution for $P(i|q)$



Interpretation of Results

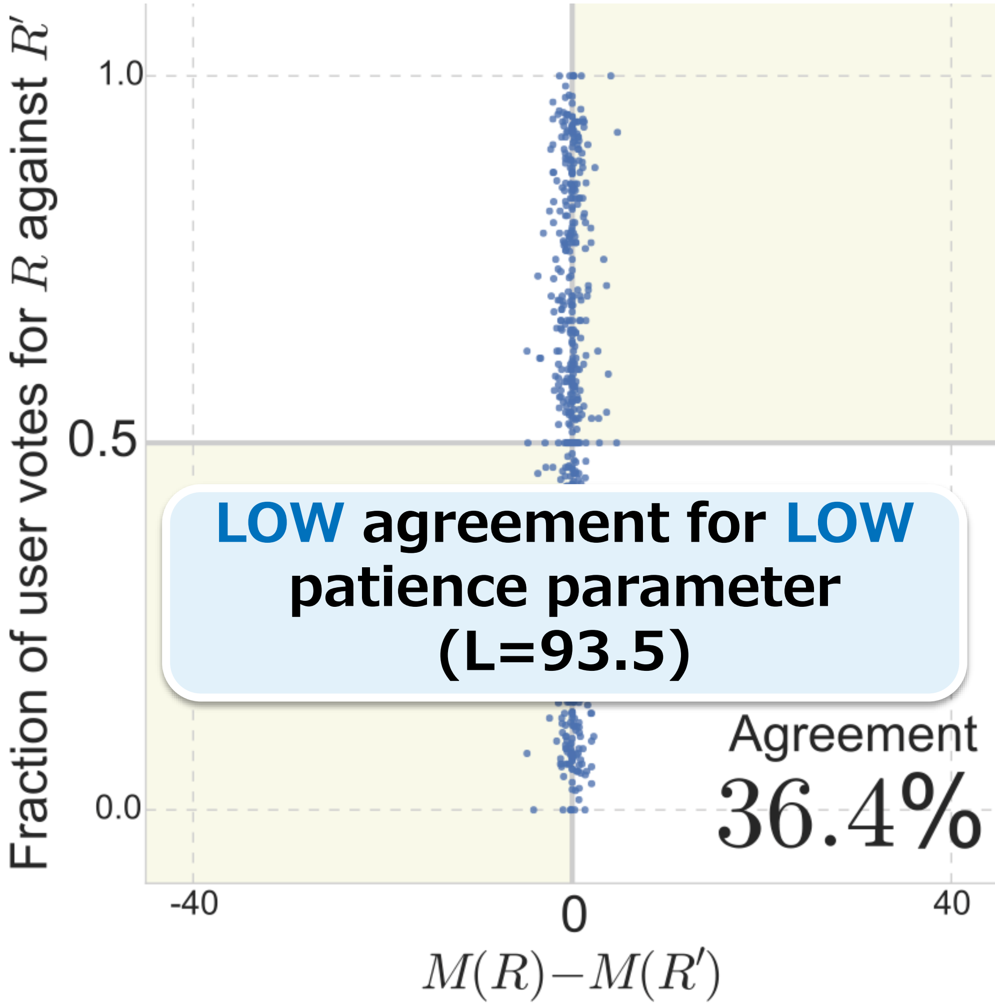
Each dot represents a pair of systems (A, B) for a particular query

$$\frac{\text{(Num. of votes for A)}}{\text{(Total num. of votes)}}$$

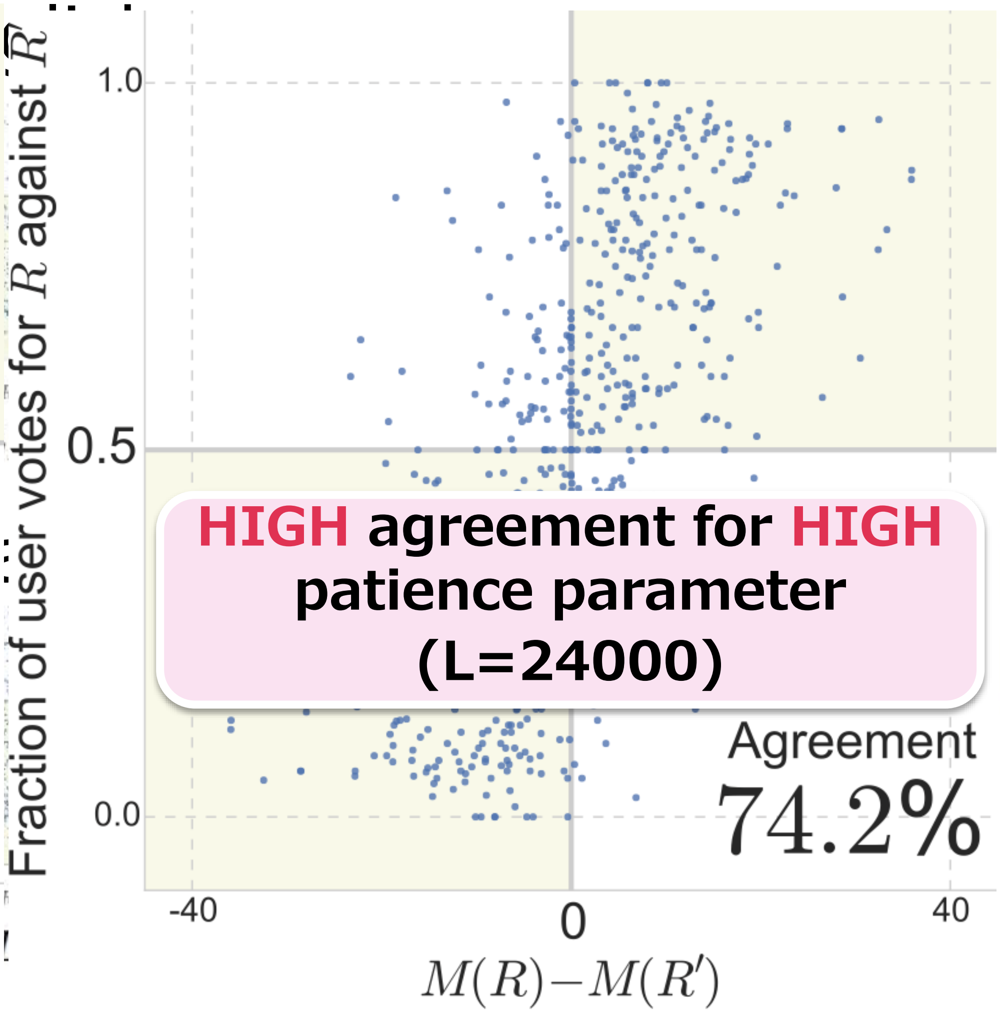


Agreement
= (#dots in Agree) / (#dots)

Experimental Results for Different Patient Parameters



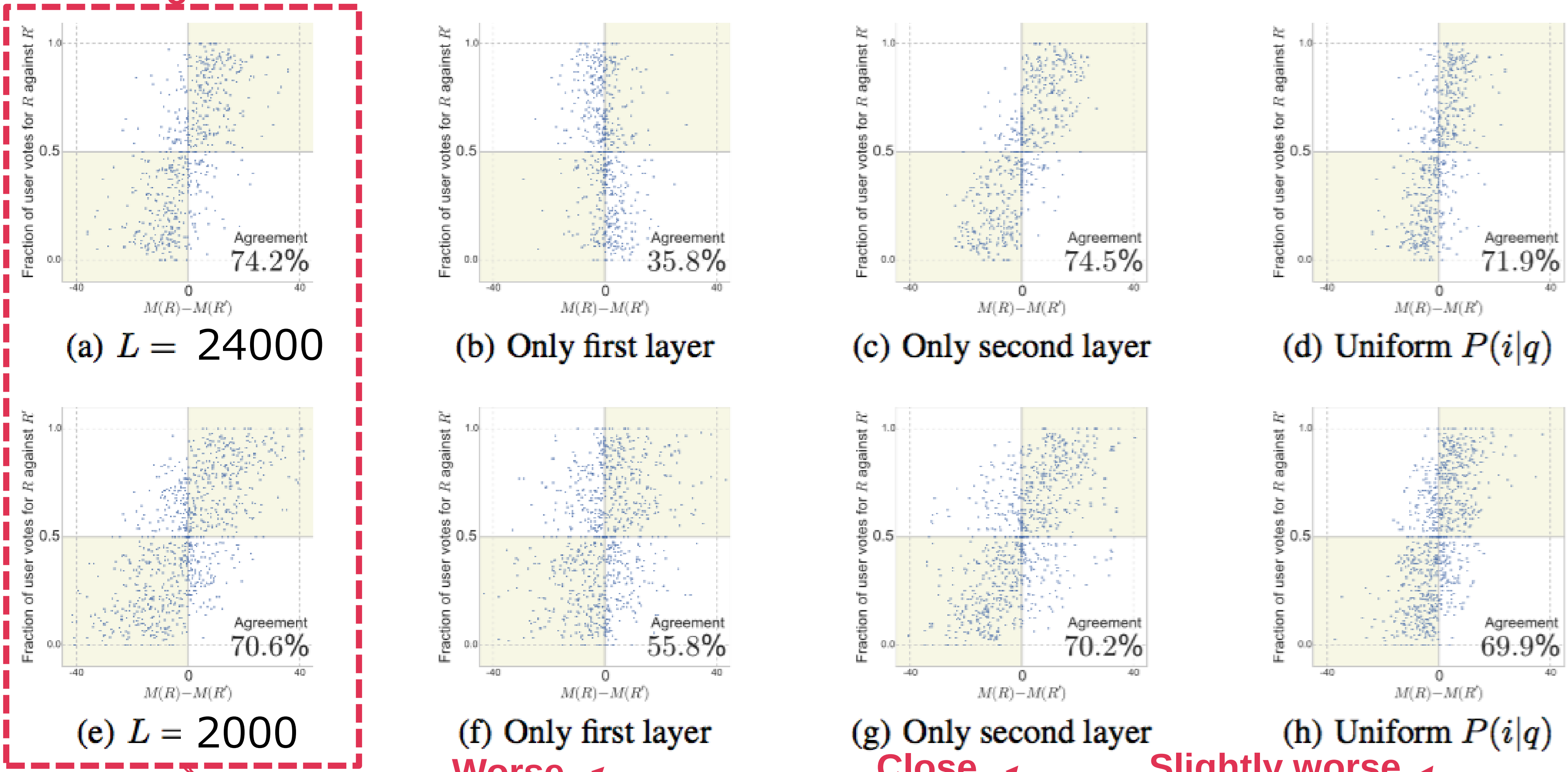
LOW agreement for LOW patience parameter (L=93.5)



HIGH agreement for HIGH patience parameter (L=24000)

Agreement is high (70-74%) for both of the languages

Experimental Results for Simple Variants of M-measure



Original

(a) $L = 24000$ (b) Only first layer (c) Only second layer (d) Uniform $P(i|q)$

(e) $L = 2000$ (f) Only first layer (g) Only second layer (h) Uniform $P(i|q)$

Worse

Close

Slightly worse

Use of the second layer and intent probability improves the agreement (but the first layer doesn't)

Why did the only 2nd layer correlate to the user pref. well?

- **Possible explanations include**
  - The quality of the second layer correlates to the quality of the whole summary
  - Users decided the quality of the summary mainly based on the second layer
    - We asked the users to look at the second layer in the assessment

Conclusions and Future Work

- **Conclusions**
  - **Proposed M-measure**
    - A special case of intent-aware U-measure for two-layered summarization
  - **Measured the agreement between M-measure and user preferences**
    - Agreement was high (70-74%)
- **Future work**
  - Error analysis
  - Address “why did the only second layer correlate to the user preferences well?”