Cyberbullying Detection Based on Category Relevance Maximization

Taisei Nitta [1]   Fumito Masui [1]   Michal Ptaszynski [1]   Yasutomo Kimura [2]   Rafal Rzepka [3]   Kenji Araki [3]

[1] Kitami Institute of Technology   [2] Otaru University of Commerce   [3] Hokkaido University

Background
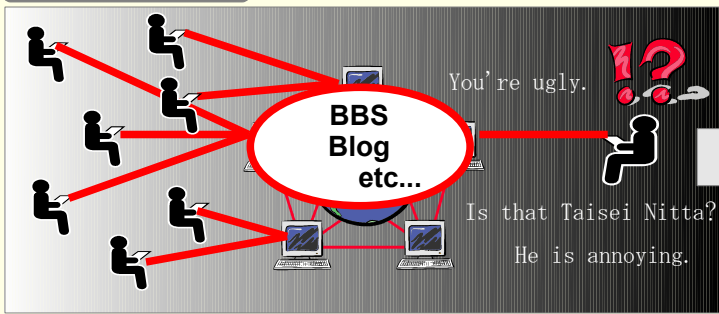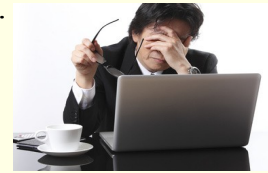

You're ugly.

Is that Taisei Nitta?

He is annoying.

Social problem of "Cyberbullying".

Net Patrol

- Monitoring activity by Parent Teacher Association (PTA)
- Request site administrator to remove harmful entries on sites.
- High cost of time and fatigue to PTA members.

Our method

Informal school Websites

"Don't get excited you ugly."

"Pet has died."

"Disappear from the world."

"Nitta is annoying."

"Could someone recommend some kind of part-time job?"

→ Detect **harmful entries** automatically →
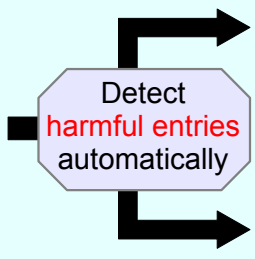
Harmful entries

"Don't get excited you ugly."

"Disappear from the world."

"Nitta is annoying."

→ Other entries

Matsuba (2011)
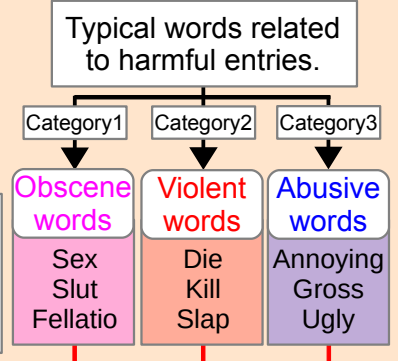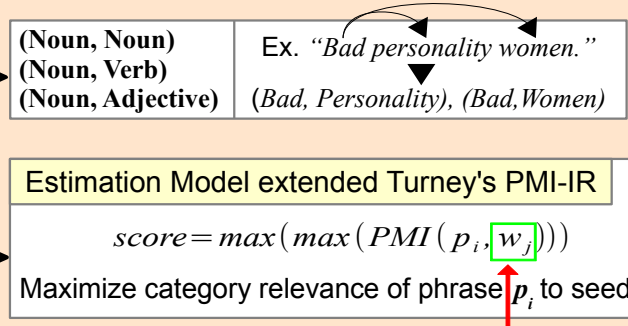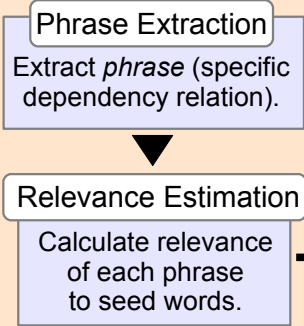
2,998 entries (Harmful entries contained 50%)

⬇

83% precision

In reality the ratio is different.

In our study,
- We propose category relevance maximization method.
- We evaluate our method's performance on data set including harmful entries of real rate.

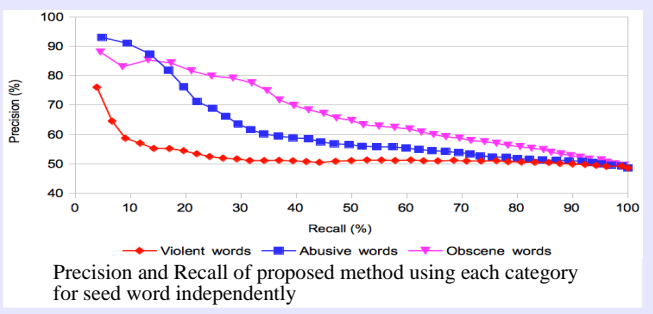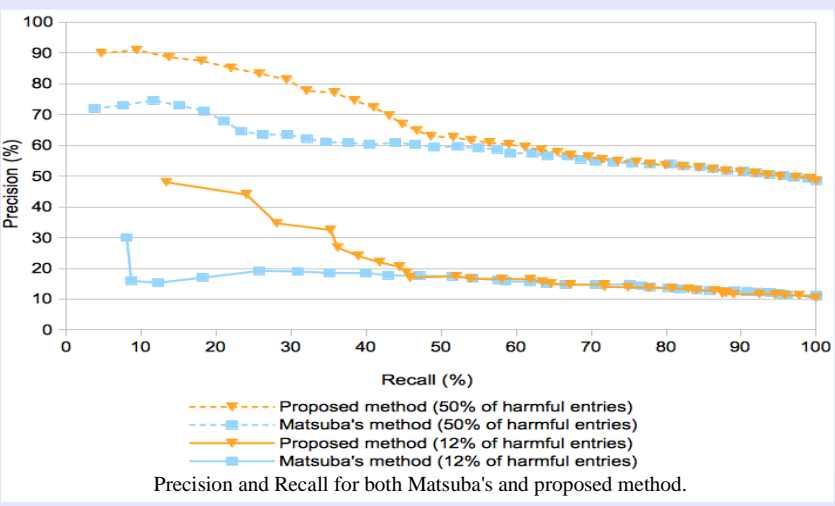## Category Relevance Maximization Method

Phrase Extraction

Extract *phrase* (specific dependency relation).

→ **(Noun, Noun)** **(Noun, Verb)** **(Noun, Adjective)**

Ex. *"Bad personality women."*

(Bad, Personality), (Bad, Women)

Relevance Estimation

Calculate relevance of each phrase to seed words.

→ **Estimation Model extended Turney's PMI-IR**

$$score = max(max(PMI(p_i, w_j)))$$

Maximize category relevance of phrase p_i to seed word w_j.

Typical words related to harmful entries.

Category1 → **Obscene words**: Sex, Slut, Fellatio

Category2 → **Violent words**: Die, Kill, Slap

Category3 → **Abusive words**: Annoying, Gross, Ugly

Experiments

◆ Test data including 500 entries. ×5
- Harmful entries contain 12%.

Actual ratio of harmful entries in informal school.

Precision and Recall for both Matsuba's and proposed method.

- ▼ Proposed method (50% of harmful entries)
- ■ Matsuba's method (50% of harmful entries)
- ▼ Proposed method (12% of harmful entries)
- ■ Matsuba's method (12% of harmful entries)

- Violent words
- Abusive words
- Obscene words

Precision and Recall of proposed method using each category for seed word independently

Discussion

- ◆ Our method's performance overcame Matsuba's method totally.
- ◆ Obscene words and abusive words contributed to perform detection of harmful entries strongly.
- ◆ Specific harmless entries were recognized as harmful.
- ◆ Personal information detection is out of range.