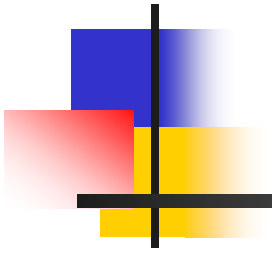


Detecting Nasty Comments from BBS Posts

Tatsuya Ishisaka

and Kazuhide Yamamoto

Nagaoka University of Technology (Japan)

Background

BBS has following comment:

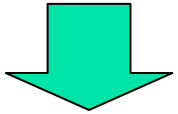
I hate you.
Everyone else hates you too.
You should just die.

Young people have been posting such comment.

In a worst-case scenario,
the victim commits suicide.

Our Goal & Approach

- Our Goal
  - The nasty comments must be managed automatically.



Detecting Nasty Comments

- Approach
  - Previous works on filtering harmful sites use harmful words as learning data. But… they are insufficient !

Because nasty comments have not only in words but also in phrases.

We also focus on nasty phrases.

Definition of Nasty Comment

- Nasty comment is defined as a sentence containing such following nasty word/phrase.

Example of the Nasty Word/phrase

・マジうざい (You are seriously annoying)

・奴らはバカな暇人野郎 (A stupid person of leisure)

Our method consists of the following four steps:

1. Building seeds dictionary of nasty words
2. Collecting nasty comments
3. Making an n-gram model
4. Detecting nasty comments

Building seeds dictionary of nasty words

- We registered 103 nasty keywords.

Example of the nasty keywords

- 死ね (You should die.)
- うざい (annoying)
- キモイ　(scumbag!)
- マスゴミ (masugomi) This is a Japanese coined word.

Collecting Nasty Comments

- We collected nasty comments automatically using seeds dictionary.
- We obtained approximately 200,000 nasty comments.

Example of the nasty comments

官僚死ねや (Bureaucrat must die.)

ゴミクズ団体はさっさと吊ってこい！ (Crap organization must perish early.)

こんなんでイチイチ騒ぐなボケカス(Keep your shirt on, chaff!)

Registered word in seeds dictionary

Making an n-gram Model 1/2

- We collected strings of words that connect with the nasty words.
- We converted nasty expression which consists of multiple words into a single word.
- We used SRILM to create a word n-gram model.

Example of the converting nasty expression

- あの バカ な マスゴミ の せい で
- あの <NASTY> の せいで

Making an n-gram Model 2/2

Example of the Nasty Words Model

0.94	<NASTY> だ な 日本 (<NASTY> da na nihon)
0.22	顔 見る と 大体 <NASTY> (kao miru to daitai <NASTY>)

Conditional probabilities
(Higher probability are nasty.)

The model has approximately 53,000 patterns.

Detecting Nasty Comments

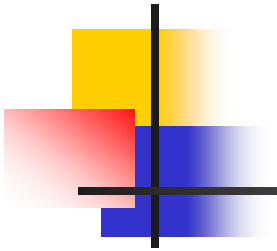
- If an input sentence includes the phrase of an n-gram model, we judge it to be a nasty comment.

The n-gram model has this phrase.

マス ゴミ の クズ どもる て，何で こう なる 事. . .

(masugomi no kuzu domoru te,　nande kou naru koto. .)

This is nasty comment !!
Because this comment contains “どもる て”.



Experiment

- Test set
  - Nasty comments: 378, Non-nasty comments: 382
  - We manually judged whether a sentence is nasty comment or non-nasty comment.
- Evaluation
  - Our method judged whether input sentences are nasty comments.

Comparative Method

- Filtering harmful information using SVM (Lee et al., 2007)
  - Feature
    - TF-IDF
    - Chi-square

    } For words
  - Training data
    - 200 to 1000 sentences

Result of F-measure

- Our method
  - The highest F-measure: 67.65 — Precision 99.74, Recall 51.17
- Comparative Method
  - The highest F-measure: 67.71 — Precision 63.15, Recall 77.81

Accuracy does not have the huge difference.
However, different type of comments were detected.

Our method:
Including nasty phrases and over-segmented nasty coined words comments
Comparative Method:
Including nasty words comments

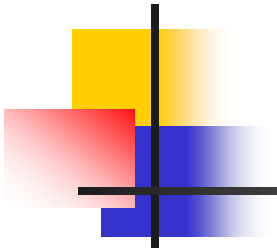
Combination Experiment

- We guess that the detection accuracy was improved by combining two methods.
- The sequential processing
  - Step1 Using our method
  - Step2 Using SVM method for nasty comments which was not detected in Step1.

Result

The highest F-measure: 72.75

Precision 61.52
Recall 89.00



Conclusion

- We have reported a method of detecting nasty comments using an n-gram from the posts on a BBS.
- Our proposed method can detect nasty comments based on nasty phrases and over-segmented words.