6th International Symposium on Advances in Science and Technology
SASTech
Malaysia | Kuala Lumpur
24 - 25 March, 2012
Scope of the Symposium
Organised by:
Khavaran Institute of Higher Education
Hosted by:
Geospatial Association of South East Asia

Thai Insult Detection System Based on Linguistic Features Analysis

Tanasanti Jirapon[1], Phokharatkul Pisit[2], Buntilov Vladimir[3], Kanoksilpatham Budsaba[4]

[1]Technology of Information System Management, Faculty of Engineering, Mahidol University, Nakorn Pathom, Thailand; [2, 3]Dept. of Computer Engineering, Faculty of Engineering, Mahidol University, Nakorn Pathom, Thailand; [4]English Department, Faculty of Arts, Silpakorn University, Nakorn Pathom 73000, Thailand

Reference Number: 6-12-11-0131
Name of the Presenter: Tanasanti, Jirapon

Abstract

Verbal insults often appear in online communities during textual communication between users. Current automatic prevention algorithms which employ regular expression techniques for word filtering tend to result in high false-positive errors. This paper presents an alternative method for detecting insults in Thai textual conversations based on the analysis of linguistic features. The performance of the presented algorithms was compared with the regular expression based algorithms, in terms of precision and recall scores. The results of the experiments showed that the inaccuracies in the employed third-party natural language processing procedures affected the performance of the proposed insult detection method. Once the problematic NLP procedures were improved, the proposed method outperforms regular expression based algorithms, showing lower false-positive error rate.

Key words: Regular Expression, Word Filter, Natural Language Processing, Online Communities, Linguistic Features

1. Introduction

Cyberspace provides people with a high degree of anonymity (Levine, 1996). People are able to communicate online without proving their identities. Some people show harassing behaviors as they hardly found themselves responsible for the acts in cyber-world. These behaviors may negatively affect the online communities. The harassing behavior focused in this study is verbal insult.

There may be several consequences from online harassments. Bond (1999) reported that some harassment resulted in law suits. A study by Campbell(2005) demonstrated negative effect of cyber bullying on child development. In some country, such as Thailand, insulting certain figures of the state is considered a criminal act. Thairath (2010) newspaper reported that some Thai website was taken down for their offensive contents. Therefore, to avoid and prevent such problems, it is important to have automatic systems which are able to detect offensive online behaviors at early stages.

This paper is organized as follows: Section 2 explores the current common solution to control online insult using Regular Expression (RE). Section 3 proposes a new method and describes the idea and usage of linguistic features in the algorithms. Section 4 describes the experiments and evaluation method. Section 5 shows results from the experiments and discusses the problems and possibilities of the algorithms.

2. Common Solution and Its Problems

A common solution to control online insults in Thai language is applying a filter based on regular expression matching. Several bulletin board systems, such as PHPBB (2008), MyBB (2009), vBulletin (2011), SMF (2011) and IP-board (Invision Power Services) and some proprietary web-board systems which were well known for Thai users, offer regular expression based word censor option. In this case, a server-based system prevents certain patterns of characters to appear on webpages. However, this solution often disrupts the communication between people as demonstrated in **Error! Reference source not found.**.

	Pattern	Word	Replace	Result
(1)	เหี้ย	เหี้ยมหาญ	***	***มหาญ
(2)	Penis	Penistone	***	***tone

Table 1 Regular expression matching which interfere with textual communication.

Error! Reference source not found. suggests that sometimes offensive words used for insult also appear in some non-insult words. An example in (1) shows Thai words. The pattern is pronounced /hia/ which is an insult. Literally, it means a water lizard. The word in focus is pronounced /hiam-harn/ which means bravery. The first part of the word /hiam/ shares the same spelling with the insult word /hia/. Regular expression based word censor replaces it as shown in the result. Another

example in (2) is the name of a town in England. Obviously, regular expression does not filter insult. It is swear filter which sometimes censors normal words.

Another example in **Error! Reference source not found.** is a news report containing the name of a sportsman, Mr. Tyson Gay. The system mistakenly replaced the word “Gay” with “Homosexual”.

The highlighted cases of regular expression failure are common in Thai language because, unlike English, Thai words in a sentence are not separated by a space. Studies, such as Sukhahuta and Smith (2001), suggested that, in order to analyze individual words, word segmentation has to be performed. Thai word segmentation process is found to be a complex task even for humans. Aroonmanakun (2002) reported that the same person may segments words inconsistently. Most current filtering methods based on regular expression do not implement word segmentation, thus censoring or replacing all matched patterns.

3. Insult detection based on linguistic features analysis

Linguistic feature is a unit of analysis for linguistic study (Kibort, 2010). It refers to the way a sound was made in speaking environment. This study focuses on text-based communication, thus the emphasis is made on writing environment. Word orders and spelling variants may affect word meanings. The underlying part-of-speech may change word meanings and word boundary in Thai is usually not clearly defined.

As was discussed before, regular expression produces many false-positive errors, thus not effective as insult detection. Wellsby et al. (2010) suggested that humans recognize insult by not just words but their contextual meanings. Therefore, a system that can recognize that meaning may be a better tool to filter out insults than regular expression.

Developing the effective insult detection system is complicated by the following problems. Firstly, some words may have several meanings. Lindén (2004) proved that the knowledge about the part-of-speech (POS) of a word helps to correctly identify the word's meaning. Furthermore, in Thai language there is no separating space between the words. Words are recognized by its meaning within given contextual information. Thus, the accuracy of a word boundary (WB) detection method is an important factor for the overall performance of the system. Both POS and WB are not utilized in RE-based implementation. The proposed method use POS and WB as a tool to solve the highlighted problems.

This study proposes an alternative for detecting insults by utilizing NLP techniques. Instead of matching patterns of characters as RE technique does, NLP enables the system to analyze the structure of each sentence and, using linguistic features such as POS and WB, to detect insults.

This work focuses on insult detection in Thai. Some pre-processing tasks are required before the actual insult detection takes place. Firstly, the analyzed text has to be segmented into sentences. Secondly, the sentences must be split into words. The third pre-processing step is to tag each word with POS. The described procedures were performed using OpenNLP package (Apache OpenNLP Development Community) configured for Thai, while ORCHID corpus (1997) was used for POS tagging. Finally, the text was structurally analyzed using tools provided by NLTK (Bird, Steven, Loper, & Klein, 2009) to establish connections between words in a sentence.

This study compares two pairs of detection algorithms. The first pair is based on Regular Expression (RE), either with only a critical word (REC), which a single insult word is given, or with a group of words (REG), which an insult word is given with some associate words. This group uses standard Java libraries to implement regular expression matching.

The second pair of the compared methods is the two versions of the proposed method based on linguistic feature analysis. The first algorithm called Insult Detection with Word Boundary (IDWB) uses semantic word boundary which separate words by its meaning. IDWB algorithm does not include parsing: the input is processed after being word-segmented. The algorithm compares each word from the input with known insult patterns. The second algorithm called Insult Detection with Part-of-Speech (IDPOS) uses both word boundary and its part-of-speech (POS). IDPOS requires input to be parsed, structurally analyzed, into tree data structure. The parsed tree is a data structure which shows hierarchical connection between each word in a sentence. The tree demonstrates dependency between words. For example, an adjective may depends upon a noun which it describes. The algorithm matches insult patterns only if the input matches both word and POS.

Insult patterns are gathered manually. The patterns are not case-sensitive because Thai do not have capital letter. The order of words shown in patterns is important. To match the patterns, the inputs must match both spelling and order of the words. Some examples are shown in **Error! Reference source not found.**.

The example of REC matches all input with a pattern of characters “insult” in it. It will match “Insult” and “Linsult?” but not match “Insulation”. The example in REG matches input with both “insult” and “associate”. For example, it matches a sentence “Insult often associate with a noun” but not matches “Insult appears alone” nor “Associate with an insult”.

The example of IDWB matches only if words, “insults” and “associate”, present in the input and are correctly segmented. For example, it matches an input which is segmented into “Insult-often-associate-with-a-noun” but not matches “Insult-often-as-so-ciate-with-a-noun”. The example of IDPOS matches only if both words and POS are in the input and correctly parsed. For example, it matches an input which has “Insult” which functions as verb if it is connected to “associate” which function as noun.

Type of pattern	Example of pattern
REC	.*insult.*
REG	.*insult.*associate.*
IDWB	(insult)(associate)
IDPOS	(verb|insult)(noun|associate)

Table 2. Example of patterns

4. Experimentation and evaluation

Experiment was conducted with three datasets. The first dataset is an article about insults in Thai. There are 453 words with 20 insults in the document. The article describes Thai insult words and demonstrates some examples. This dataset is used to determine if proposed algorithms could identify insults shown in examples. Words description should not be match.

The second one simulates exchange between persons with noticeable insults. There are 496 words with 10 insults in the document. This dataset is used to test proposed algorithms performance within textual communication environment. There are some numbers and symbols appear in the text. The article, unlike the first dataset, was not formally organized.

Some performance problems are observed during experiment with the second dataset. The problem is hypothesized to be cascaded errors from pre-processing. Therefore, the last dataset is used to confirm it. The third dataset is similar to the second except the words are space-separated to aid segmentation process.

The goal of the evaluation experiments is to test if the proposed insult detection method is able to effectively detect insults and to compare its performance with classical algorithms. Each algorithm was tested on the datasets and returned number of insults it detected. The methods were evaluated using Precision and Recall scores demonstrated by Baeza-Yates and Ribeiro-Neto (1999). Both scores were weighted to F-Score, the harmonic means of two values, thus, become evaluation score shown in equation 1.

$$Score = 2 * \frac{\frac{t}{t+f} * \frac{t}{t+e}}{\frac{t}{t+f} + \frac{t}{t+e}} \quad (1)$$

where, t represents the number of correct results, f represents the number of incorrect results. e is the number of correct results which the algorithm failed to retrieve.

5. Result and Analysis

The experiment with the first dataset showed some accuracy issues. As expected, REC and REG detected all insults but produced many false-positive errors. IDWB and IDPOS were not as accurate in detecting insults but showed less false-positive errors. F-score suggested that using REG was slightly better than using WB without POS. REC showed lowest score while IDPOS had the highest score.

Technique	Name	t	f	e	Precision	Recall	F-Score
Regular Expression	REC	20	10	0	0.67	1	0.80
Regular Expression	REG	20	4	0	0.83	1	0.91
Linguistic Features Analysis	IDWB	19	3	1	0.86	0.95	0.90
Linguistic Features Analysis	IDPOS	19	1	1	0.95	0.95	0.95

Table 3. Results and scores from the first dataset

The experiment with the second dataset showed a problem which severely affected detection performance of linguistic features analysis algorithms. REC and REG detected all insults but also high false-positive errors. IDWB and IDPOS missed many insults. However, the final F-Score of IDPOS was slightly higher than both REC and REG. Similarly to the result from the first dataset, IDWB got higher score than REC but slightly lower than REG.

Technique	Name	t	f	e	Precision	Recall	F-Score
Regular Expression	REC	10	13	0	0.43	1	0.61
Regular Expression	REG	10	11	0	0.48	1	0.65
Linguistic Features Analysis	IDWB	5	1	5	0.83	0.5	0.63
Linguistic Features Analysis	IDPOS	5	0	5	1	0.5	0.66

Table 4. Results and score from the second dataset

The result from second dataset suggested series of cascade errors from NLP. For example, some sentences were not correctly segmented, which led to wrong word segmentation and inaccurately parsed sentences. Finally this negatively affected overall detection performance.

The inaccuracy of pre-processing steps was identified and the experiments continued with manual tweak of NLP procedures, particularly of word segmentation. The third dataset was created from the second dataset by manually guided word segmentation. The accuracy of the NLP pre-processing was improved. Despite improvement in NLP tasks, accuracy of REC and REG did not improved. In fact, REC got lower score because, in the experiment with the second dataset, sometimes NLP pre-processing incorrectly segmented two sentences into one sentence thus make one positive instead of two. REG did not affected by such problem because it requires both insult words and associate words, such as a preposition, in particular order. IDWB and IDPOS got higher scores than both REC and REG. IDPOS got slightly higher score than IDWB. This showed importance of POS in detection.

Technique	Name	t	f	e	Precision	Recall	F-Score
Regular Expression	REC	10	15	0	0.4	1	0.57
Regular Expression	REG	10	11	0	0.48	1	0.65
Linguistic Features Analysis	IDWB	10	1	0	0.91	1	0.95
Linguistic Features Analysis	IDPOS	10	0	0	1	1	1

Table 5. Results and score from the third dataset

Average F-Scores from each dataset suggested that linguistic features analysis algorithms were able to filter insults to a reasonable degree, slightly better than RE-based algorithms in all datasets.

Dataset	Regular expression			Linguistic features analysis		
	Precision	Recall	F-Score	Precision	Recall	F-Score
1st dataset	0.750	1.000	0.855	0.905	0.950	0.925
2nd dataset	0.455	1.000	0.630	0.915	0.500	0.645
3rd dataset	0.440	1.000	0.610	0.955	1.000	0.975

Table 6. Average scores of different algorithms by datasets

The scores showed that using inaccurate word-segmentation procedures significantly deteriorate the performance of the proposed method. However, even in this case linguistic features analysis algorithms produced less false-positive errors than regular expression based algorithms.

Further studies which improve performance of NLP could positively affect insult detection method proposed in this study. Insult patterns could be done automatically, using supervised learning algorithms, to filter new insults. The algorithms should be tested with a larger dataset in real-time communication to measure their performance. Finally, other linguistic features, used in combination with POS and WB, might improve insult detection.

6. Conclusions

A new insult detecting method which utilizes linguistic features of text is proposed. This paper explores some problems posted by insult in online communities. Regular expression technique is a common tool to implement insult detection, but it introduces many false-positive errors. The carried out experiments compared regular expression-based algorithms with two versions of the proposed technique. The experiments used simulated Thai textual data which contain insults. RE-based algorithms showed many false-positive errors but matched most insults. The proposed methods produced less false-positive errors but are severely affected by the cascaded errors from NLP pre-processing procedures (e.g. word segmentation). The experiments suggested that once the pre-processing steps are improved the proposed methods offered superior accuracy over the conventional methods.

Acknowledgements

I would like to express great appreciation to my family, who support me in pursuing for knowledge, and my advisors: Dr Pisit Phokharatkul, Dr. Vladimir Buntilov and Dr. Budsaba Kanoksilpatham, who make this work possible. This work is supported by the 60th Year Supreme Reign of his Majesty the King Bhumibol Adulyadej Scholarship of the Faculty of Graduate Studies, Mahidol University.

References

Apache OpenNLP Development Community. (n.d.). *Apache OpenNLP Developer Documentation.* Retrieved September 4, 2011, from Apache OpenNLP Developer Documentation: http://incubator.apache.org/opennlp/documentation/manual/opennlp.html

Aroonmanakun, W. (2002). Collocation and Thai Words Segmentation. *Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop* (pp. 68-75). Pathumthani: Sirindhorn International Institute of Technology.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* New York: Addison Wesley.

Bird, Steven, Loper, E., & Klein, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

Bond, R. (1999). Links, Frames, Meta-tags and Trolls. *International Review of Law, Computers & Technology*, 317-323.

Campbell, M. A. (2005). Cyber bullying: An old problem in a new guise? *Australian Journal of Guidance and Counselling*, 68-76.

Charoenporn, T. S. (1997). Building A Large Thai Text Corpus - Part-Of-Speech Tagged Corpus: ORCHID. *Proceedings of NLPRS.*

Invision Power Services. (n.d.). *Post Content: Bad Word Filters.* Retrieved September 3, 2011, from Invision Power Services: http://community.invisionpower.com/resources/documentation/index.html/_/documentation/administrator-control-panel/look-and-feel/post-content-bad-word-filters-r307

Kibort, A. (2010, July 17). *Feature Inventory.* Retrieved September 4, 2011, from Grammatical Features: http://www.grammaticalfeatures.net/inventory.html

Levine, N. (1996). Establishing Legal Accountability for Anonymous Communication in cyberspace. *Columbia Law Review*, 1526-1528.

Lindén, K. (2004). Evaluation of Linguistic Features for Word Sense. Disambiguation with Self-Organized Document Maps. *Computers and the Humanities*, 417-435.

MyBB Group. (2009, January 27). *Admin CP.* Retrieved September 3, 2011, from MyBB Wiki: http://wiki.mybb.com/index.php/Admin_CP

Online News Team. (2010, April 11). *Thairath Online News Report.* Retrieved September 4, 2011, from Thairath Online: http://www.thairath.co.th/content/tech/76263

phpBB Group. (2008). *Documentation 3.4. Posting Settings.* Retrieved September 3, 2011, from phpBB 3.0 Olympus Documentation: http://www.phpbb.com/support/documentation/3.0/adminguide/acp_posting.php

Simple Machines. (2011, June 15). *Using SMF as an administrator: Ban List.* Retrieved September 2011, 3, from Simple Machines Online Manual: http://wiki.simplemachines.org/smf/Ban_list

Sukhahuta, R., & Smith, D. (2001). Information Extraction Strategies for Thai Documents. *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, 153-172.

vBulletin Solutions. (2011). *Censorship Options* . Retrieved September 3, 2011, from vBulletin Manual: https://www.vbulletin.com/docs/html/vboptions_group_censor

Wellsby, M., Siakaluk, P. D., Pexman, P. M., & Owen, W. J. (2010). Some insults are easier to detect: the embodied insult detection effect. *Frontiers in Psychology*, 198.1-198.13.