Journal of ICT, 12, 2013, pp: 147–159

NORMALIZATION OF NOISY TEXTS IN MALAYSIAN ONLINE REVIEWS

Norlela Samsudin[1], Mazidah Puteh[2], Abdul Razak Hamdan[3] and Mohd Zakree Ahmad Nazri[4]

[1&2]Faculty of Computer and Mathematical Science,
Universiti Teknologi MARA Terengganu,
Dungun, 23000, Terengganu, Malaysia

[3&4]Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
43600, Bangi, Selangor, Malaysia

Corresponding author: norlela@tganu.uitm.edu.my[1]

ABSTRACT

The process of gathering useful information from online messages has increased as more and more people use the Internet and other online applications such as Facebook and Twitter to communicate with each other. One of the problems in processing online messages is the high number of noisy texts that exist in these messages. Few studies have shown that the noisy texts decreased the result of text mining activities. On the other hand, very few works have investigated on the patterns of noisy texts that are created by Malaysians. In this study, a common noisy terms list and an artificial abbreviations list were created using specific rules and were utilized to select candidates of correct words for a noisy term. Later, the correct term was selected based on a bi-gram words index. The experiments used online messages that were created by the Malaysians. The result shows that normalization of noisy texts using artificial abbreviations list compliments the use of common noisy texts list.

Keywords: Noisy texts; normalization of noisy texts; artificial abbreviation

INTRODUCTION

The advancement of Internet technology causes a mass collection of online documents from applications such as e-forums, blogs, Facebook and Twitter.

The online social media allow the users to communicate with each other in an informal environment. Therefore, the online documents are filled with out of vocabulary (OOV) terms or noisy texts and do not follow the usual structure of a language. Knoblock, Lopresti, Roy and Subramaniam (2007) define noisy text as "any kind of difference between the surface form of a coded representation of the text and the *intended*, *correct*, or *original* text". Despite being noisy, online created documents contain important information such as opinions about a particular product, service or political figure. Other than that, customers often give feedbacks or comments about an organization using online facility. Mining the online documents may reveal interesting information for the survival of a company. Frequently Asked Question (FAQ) is another application that received input from the customer via the online application. Unfortunately, the noisy texts that exist in online messages lead to inaccurate information in text processing activities. Therefore, processing of online documents is necessary before any information gathering activities from online created messages is executed. The following is an example of a typical e-forum entry that is created by Malaysian:

> **"***budak kecik ni asyik sangat tengok 7 petala cinta. br lps tgk citer ni (-____-)……. best citer ni!!! bc komen2 kt sini...yg mana lost2 boleh faham balik... http://asdkfj.kasdf.dfjk.my ".*

This message is filled with incorrect sentence structure, improper casing, incorrect punctuation, misspelled words, mixed of terms from different languages and creative use of emoticon. Work by Samsudin, Puteh and Hamdan (2011) and Dey and Haque (2009) showed that the occurrence of noisy texts reduced the accuracy value of opinion mining processing. Similarly, Vinciarelli (2004) concluded that noisy text also affects text mining activities. Other than that, experiments by Tang, Li, Cao, and Tang (2005) also concluded that the terms extraction from electronic mails was improved by 35% to 45% after the emails had been cleaned from noisy terms.

Normalization of noisy texts in previous researchers uses resources mainly from English language such as:

1) a standard parser which is used by Clark (2003), Foster, Wagner, and Genabith (2008), Jing, Lopresti, and Shih (2003);
2) resources from Word Wide Web in Wong, Liu, and Bennamoun (2006);
3) English dictionaries in Wong, Leu, and Bennamoun (2006), Dey and Haque (2008) or
4) specific domain dictionary used by Kothari, Negi, Faruquie, Chakaravarthy, and Subramaniam (2009).

Unfortunately, there is no such reference that is available for the Malay language. In addition, most previous works try to solve noisy terms involving words created from its phonetic sound such as 'cu', 2u, 2morrow, l8, or *lol*. Malaysians rarely use these terms. This study shows that the top five noisy terms that are commonly used by Malaysians in online documents are *tu (itu), yg (yang), ni (ini), tak (tidak)* and *x (tidak).* The shorter version of a term or abbreviation is used in order to reduce key punching (especially when a mobile hardware is used to create the message) and to speed up the communication process. This project studied the pattern of abbreviations that Malaysians used in online media and created artificial abbreviations list to improve the normalization process of noisy texts. In addition to that, a list of common noisy texts that Malaysians normally used in online message was also created and used in the normalization process.

BACKGROUND

Kobus (2008) identified three metaphors in cleaning noisy texts i.e. spell checking metaphor, translation metaphor and speech recognition metaphor. Spell checking metaphor assumes all out of dictionary words as noisy terms and need to be corrected. This technique normally uses a specific dictionary to identify a noisy term. Most works in normalization of noisy texts adopt this metaphor such as work by Toutanova and Moore (2002), Wong, Leu et al. (2006), Choudhury et al. (2007) and Cook and Stevenson (2009). Nevertheless, this method does not consider the context where the term is used. The second metaphor assumes texts with noisy term as another language and uses a specific dictionary to translate these texts into the correct texts. The researchers normally use statistic techniques to solve the problem such as phrase-based statistical model by Aw, Zhang, Xiao and Su (2006) and Hidden Markov Model in Choudhury et al. (2007) and Acharyya, Negi, Subramaniam, and Roy (2008). The last metaphor is based on works to convert speech notation into texts. Users of online communication normally communicate in an informal manner. The use of texts which imitates the phonetic sound of a word, such as *fon* (*phone*), *2nite* (*tonight*) or *cite* (*cerita*), is common in online communication. This method uses predefined codes that translate phonetic sound spelling to written texts spelling based of specific rules (Kobus, 2008).

One of the trends in online messages is using shortened words in the form of acronym or abbreviation. Acronym is a word that is formed by combining the initial letters from a group of words such as UUM (Universiti Utara

Malaysia), AF (Akademi Fantasia) and *lol* (*laugh out loud*). On the other hand, abbreviation is a shortened form of a word such as *gd* (*good*), *bst* (*best*) and *kg* (*kampong*). Constrain of a device due to the use of mobile phone as a medium of communication and constrain of time cause online users to shorten the spelling of texts in online messages. Several trends on how Malaysians shorten the Malay terms have been identified in Hussin (2009) and Pustaka (2008). This paper investigates the used of common noisy terms list and artificial abbreviations list to normalize noisy texts. To the knowledge of the writer, this work is the first attempt to normalized online messages that are written by Malaysians.

METHODOLOGY

Preparation of data

The experiment requires a collection of online messages created by Malaysians. In order to create this collection, 5000 e-forum entries, 5000 Twitter messages and 5000 Facebook messages believed to be created by Malaysians were manually extracted. As shown in Figure 1, the following lists were created from these messages:

a) A list of noisy terms that occur more than three times in these documents. About 4000 noisy terms have been identified and manually translated. This list is known as **NTTranslate**.
b) A list of all correctly spelled words other than proper names. Items from this list are merged with translation of noisy text from list (a). A total of 10550 words are listed. This list is named as **CommonWords**.
c) The contents of corrected spelled words from (b) are merged with a list of Malay words taken from a digital dictionary. This list is known as **CorrectWords** list and is used in the project to identify an out of vocabulary (OOV) term.
d) The online documents were semi-automatically translated and verified. A list that records the frequency of bi-gram words in the corpus was created and used to select the most suitable term as a translation for a noisy text. This list is known as **Bi-Gram Index**.

Another 100 online messages were extracted as testing data. Noisy texts were tagged and translated manually. Other terms were tagged as correct word, numbers, icon, link and symbol. These data were used to check the effectiveness of normalization processes in this study.

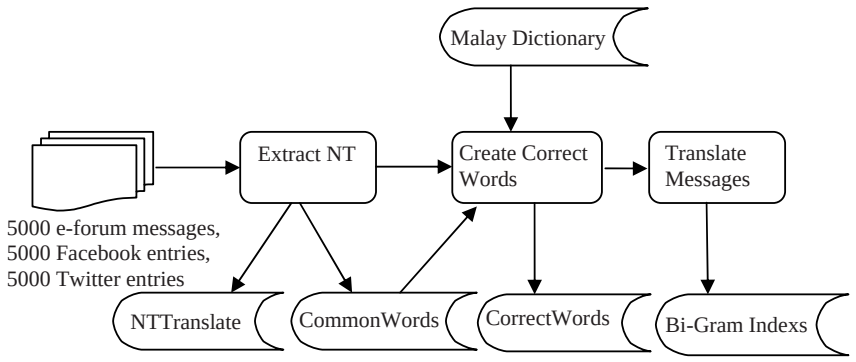

Figure 1. Processing of online messages corpus.

Generating artificial abbreviations list

A Malay term is made of several syllables. A syllable is the smallest unit of a speech sound. Normally it is made from several combinations of a vowel and consonants. For example word 'kucing' is a combination of two syllables i.e 'ku' and 'cing'. In addition to the normal consonant character, the Malay language also adopts group consonants i.e. *gh, kh, ny, ng, sy*. The rules in creating artificial abbreviation manipulate the characters and syllable of a particular word. In 2008, a guideline in creating SMS abbreviation in Malay Language was published by Dewan Bahasa & Pustaka. Adopting these rules and observation on the abbreviation pattern of the top 200 noisy texts, a list of artificial noisy texts is created. Rules that are related to manipulation of characters are:

1. Remove all vowels such as in *sklh (sekolah)* and *slr (seluar)*
2. Use the first character and the last character if either of them is not a vowel such as *yg (yang)* and *kg (kampong).*
3. Replace the last character with the character 'e' if it is an 'a' such as *ape (apa)* and *berape (berapa).*
4. Add character 'k' to the end of the word if the word is ended with character 'a' such as *bapak (bapa)* and *mintak (minta).*
5. Drop the first vowel if the word starts with a consonant such as *sapa (siapa), slalu (selalu)*
6. Drop the last vowel if the last character is not a vowel such *ank (anak)* and *ingt (ingat).*
7. Using the first and the last character such as *pi (pergi)* and *dn (dan).*

8. If the term ends with ‘ar’, replace it with the character ‘o’ such as *sabo (sabar)* and *terbako (terbakar)*
9. If the term starts with ‘ha’, drop the character ‘h’ such as *antu (hantu)* and *ari (hari).*
10. Using a character in replacement to a word with similar phonetic sounds is also common. The following abbreviations are also added in the list: *w (why), x (tidak), n (dan), g (pergi), s (as), d (di), k (ok), u (you), t (nanti)*

The following rules manipulate the syllables of a word.

1. Use the first syllable such as *sem (semester).*
2. Use the last syllable such as *mak (emak)* or *ngan (dengan).* If the new last syllable ends with an ‘a’, replace it with ‘e’ such as *je (sahaja)* or *te (kita).* In addition to that, if the character ends with an ‘a’, add character ‘k’ such as *gak (jugak);*
3. Use the first character of each syllable in a word such as *spt (seperti).* If the syllable starts with a group of consonant, the second character will be used such as *tgk (tengok);*

In addition rules that are listed previously, the following rules that manipulate the syllables and the characters are also adopted.

1. Use the first character of each syllable + the last character (if the word is a consonant) such as *byk (banyak)* and *tgh (tengah).*
2. Use the first character and the last syllable such as *bleh (boleh)* or *bru (baru).* If the new term ends with an ‘a’, replace it with the character ‘e’ such as *bpe (berapa)* and *mne (mana).*
3. Use the last syllable but replace the first character of the last syllable with the first character of the word such as *tak (tidak)* and *tgok (tengok).*

Using **CommonWords** list, about 80,000 artificial noisy texts were created and named as **Artificial Abbreviation** list.

Normalization process

Three experiments were conducted in this project. The first experiment is considered as the based experiment, where normalization of noisy text was executed using the common noisy texts translation. If more than one translation were identified, the correct term was selected at random. Figure 2 illustrates the process.

Kukich (1992) suggests three stages in the normalization process of noisy texts named Detect Noisy Terms, Identify Candidates and Choose Translation a. In order to identify a noisy term, a word is compared to a list of dictionary which contains correct words. All words that are not in the dictionary are considered as noisy terms. The next step identifies the candidates of correct words using a list of artificial abbreviation which has been created using rules that have been explained in the previous section. The last step identifies the correct term based on the context where the word is used. This is done by comparing the occurrences of the previous word. These steps made up the second experiments as illustrated in Figure 3.

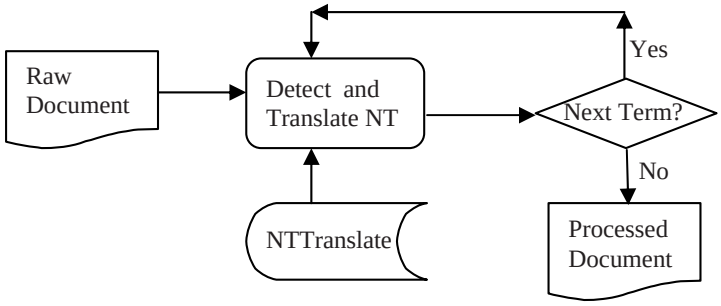

Figure 2. Using common NT list.

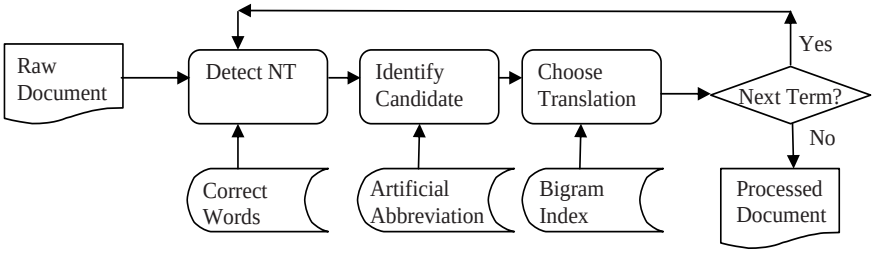

Figure 3. Normalization of noisy text using artificial abbreviation.

In the third experiment, in addition to artificial abbreviation, the common noisy terms list is added as one of the references in identifying correct term candidates as illustrated in Figure 4.

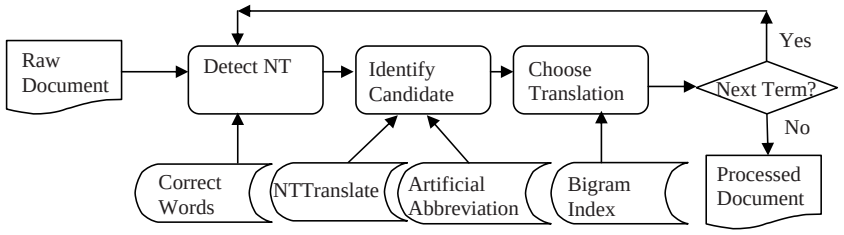

Figure 4. Normalization of noisy text using common noisy text and artificial abbreviation list.

RESULT AND DISCUSSION

The purpose of this study is to check whether artificial abbreviation lists and common noisy text translation can improve the process to 'clean' noisy terms in an online media message that were created by Malaysians. 100 online messages that stated opinions about a particular movie had been extracted from various e-forum, Facebook entries and Tweeter messages. These messages contain between 11 and 170 words with an average of 60 words per message. On average, 15 noisy texts were identified manually in every message. Surprisingly, the system identified an average of 17 noisy texts in every message. This is due to the use of English words that were not listed in **CorrectWords** list. This list was created using common words in 15,000 online messages and a Malay dictionary. Therefore, words such as *predictable*, *private* and *characters* were considered as noisy terms since these words did not exist in **CorrectWords** list. In the researchers' opinion, the English terms that exist in **CommonWords** list are enough to identify the common English words used by Malaysians in online messages. Unfortunately, that is not the case as shown by an increase of 2% in noisy texts identified by the system. Other than that, a proper noun, such as the name of a person or a movie that was spelled without using an upper case letter as the first character, was also considered as a noisy term. Therefore, the number of noisy texts that was identified in every experiment was higher than the number of noisy texts that was identified manually. Correctly identified noisy text is noisy text that was correctly identified and translated as identified and translated in the manual process. Incorrect identified noisy text is a word that was not considered as noisy text in the manual process or a word but was identified as noisy text and translated wrongly. Table 1 shows the average percentage of correctly identified noisy texts and the average percentage of incorrectly identified noisy texts that were captured at each experiment.

Table 1

Results of Experiments

	NTTranslate	Artificial Abbreviation	NTTranslate + Artificial Abbreviation
Correctly Identified NT	70%	42%	76%
Incorrect Identified NT	40%	58%	34%

The result of the experiment shows that 70% of noisy texts that were identified in the messages may be corrected using the common noisy texts list. On the other hand, only 42% of noisy texts can be corrected using the list of artificial abbreviation alone. Nevertheless, the result improved when both lists were used. **NTTranslate** is a list of manually noisy text translation which is extracted from 15000 online media messages created by Malaysians. Therefore, common noisy texts were captured in this list and produced a better result in the normalization of noisy texts as compared to using artificial noisy text list alone. Unfortunately, using only the common noisy terms list had several set- backs. It failed to capture the relation between a word and its previous word. Neither can it identify other creative short forms of a word that were not commonly used. In addition, processing of noisy terms that consist of a number is limited to the existence of the word in the common noisy text list. This problem was tackled when artificial abbreviation list is used. Even though the artificial abbreviation list solves the above problems, it cannot recognize noisy terms that use phonetic similarity such as *2CU (to see you), pilem (filem), citer (cerita)* and *siyes (say yes).* Other than phonetic sound, the approach in this study also ignores the following type of noisy texts.

a) Abbreviation that is made from a combination of two or more word such as dorg (dia orang) and pastu (selepas itu).
b) Identification of proper names such as the name of a person or the name of a country. Currently, the algorithm assumes words that start with a capital letter as proper name and hence, they will not be processed. On the other hand, a proper name that starts with lower case letter is assumed as noisy text.
c) Double meaning word. For example, sapa is a root word and is being used in words such as menyapa or disapa. This word is considered as a correct word and exists in a dictionary. Nevertheless, when it is used in in a different context such as “kubur sapa ni? “ where the word sapa is considered as a noisy term. The correct word is siapa. This situation was not identified and corrected in the normalization process.

(d) Slang words such as ma, je, jee, le, bah, gezek and lu. Other than these words, terms that indicate expressions, such as augh, err, haha, and hehehehe are also ignored. These words are considered as correctly identified noisy terms.

Other reasons for incorrect translation are:

- Typing errors such as the word **tima** in the phrase **'tima aku tengok'** is supposed to be **'time'**. Since the word **'tima'** is not considered as common noisy term, it is not listed in **NTTranslate**, but the term is listed in abbreviation list as the short term for word **'terima'**.
- Noisy texts from unlisted word in digital dictionary such as **sgtle** which means **sangatlah**. This word occurs due to the additional suffix added by the users.
- **Creative words that the users used which are out of norm and so do follow the usual pattern of noisy terms creation such as ritu (hari itu),pes (peace)** and asik (asyik)

CONCLUSION

This study showed that common noisy texts list and artificial abbreviation list were effective in the normalization of noisy terms where Malaysian online messages were involved. Both lists are the main contributions of this study. The common noisy texts list is a list of noisy texts that occurred three times or more in 15,000 online messages created by Malaysians. Nevertheless, noisy terms that are used by the online users varies based on the environment or domain of the subject. Therefore, the artificial abbreviation list complements the common noisy texts list and produced a better result in the normalization process. The artificial abbreviation list is created by projecting noisy terms that the users may use based on several common patterns of short forms observed by the researchers.

At the end of the study, the researchers believed that incorporating other modules could improve the result of noisy text normalization. Among them are:

1) using English dictionary in addition to the Malay dictionary to identify OOV words;
2) incorporating a technique to check noisy term as the result of using suffix and prefix on the Malay words;

3) incorporating a technique to solve words that follow the phonetic sound instead of how it is spelled; and
4) incorporating a list of slang words and words that express expressions such as *arg*, *oh*, *zzzz*, and *hurg*.

The impact of using these modules is possible enhancements on the research in the future. As more and more people use the Internet or other online applications to communicate with each other, the need to process online text messages will also increase. The noisy texts that are incorporated in these messages need to be normalized so that other text processing applications such as Q & A, customer services, classification and information retrieval, may produce useful and accurate information. The common noisy text list and the artificial abbreviation list are two references that may be utilized in noisy texts normalization process for messages created by Malaysians.

ACKNOWLEDGEMENT

This research is supported by the Fundamental Research Grant Scheme (FRGS) under the ninth Malaysia Plan (RMK-9), Ministry of Higher Education (MOHE) Malaysia. The grant number is 600-RMI/ST/FRGS 5/3/Fst (208/2010).

REFERENCES

Acharyya, S., Negi, S., Subramaniam, L. V., & Roy, S. (2008). *Unsupervised learning of multilingual short message service (SMS) dialect from noisy examples*. Paper presented at the Second Workshop on Analytics for Noisy Unstructured Text Data, Singapore.

Aw, A., Zhang, M., Xiao, J., & Su, J. (2006). *A phrase-based statistical model for SMS text normalization*. Paper at the COLING/ACL on Main Conference Poster Sessions, Sydney, Australia.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., & Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition, 10*(3), 157-174. doi: 10.1007/s10032-007-0054-0

Clark, A. (2003). *Pre-processing very noisy text*. Paper presented the Workshop Shallow Processing at Large, Corpora, Lancester.

Cook, P., & Stevenson, S. (2009). *An unsupervised model for text message normalization*. Paper presented at the Workshop on Computational Approaches to Linguistic Creativity, Boulder, Colorado.

Dey, L., & Haque, S. K. M. (2008). *Opinion mining from noisy text data*. Paper presented at the Second Workshop on Analytics for Noisy Unstructured Text Data. Singapore.

Dey, L., & Haque, S. K. M. (2009). *Studying the effects of noisy text on text mining applications*. Paper presented at the Third Workshop on Analytics for Noisy Unstructured Text Data, Barcelona, Spain.

Foster, J., Wagner, J., & Genabith, J. V. (2008). *Adapting a WSJ-trained parser to grammatically noisy text*. Paper presented at the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, Ohio.

Hussin, S. (2009). *Bahasa SMS*. Retrieved from http://supyanhussin.wordpress.com/2009/07/11/bahasa-sms/

Jing, H., Lopresti, D., & Shih, C. (2003). *Summarization of noisy documents: a pilot study*. Paper presented at the HLT-NAACL 03 on Text summarization workshop - Volume 5.

Kobus, C., Yvon, F., & Damnati, G. (2008). *Normalizing SMS: Are two methaphors better than one?* Paper presented at the 22nd International Conference on Computational Linguistics, Manchester.

Kothari, G., Negi, S., Faruquie, T. A., Chakaravarthy, V. T., & Subramaniam, L. V. (2009). *SMS based interface for FAQ retrieval*. Paper presented at the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, Suntec, Singapore.

Dewan Bahasa dan Pustaka, (2008). *Panduan Singkatan Khidmat Pesanan Ringkas*. Retrieved from http://www.dbp.gov.my/khidmatsms.pdf

Samsudin, N., Puteh, M., & Hamdan, A. R. (2011). *Bess or xbest: Mining the Malaysian online reviews*. Paper presented at the 3rd Conference on Data Mining and Optimization (DMO).

Tang, J., Li, H., Cao, Y., & Tang, Z. (2005). *Email data cleaning*. Paper at the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA.

Toutanova, K., & Moore, R. C. (2002). *Pronounciation modeling for improved spelling correction*. Paper presented at the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania.

Vinciarelli, A. (2004, 23-26 Aug). *Noisy text categorization.* Paper presented at the 17th International Conference on Pattern Recognition, ICPR.

Wong, W., Leu, W., & Bennamoun, M. (2006). *Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text.* Paper presented at the Australasian Data Mining Conference, Sydney.

Wong, W., Liu, W., & Bennamoun, M. (2006). *Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text*. Paper presented at the Fifth Australasian Conference on Data Mining and Analystics - Volume 61, Sydney, Australia.