

Proposal: A Hybrid Dictionary Modelling Approach for Malay Tweet Normalization

Nor Azlizawati Binti Muhamad, Norisma Idris and Mohammad Arshi Saloot

University of Malaya, 50603, Kuala Lumpur, Malaysia.

azlizamuhamad@gmail.com

Abstract. Malay Twitter message presents a special deviation from the original language. Malay Tweet widely used currently by Twitter users, especially at Malaya archipelago. Thus, it is important to make a normalization system which can translated Malay Tweet language into the standard Malay language. Some researchers have conducted in natural language processing which mainly focuses on normalizing English Twitter messages, while few studies have been done for normalize Malay Tweets. This paper proposes an approach to normalize Malay Twitter messages based on hybrid dictionary modelling methods. This approach normalizes noisy Malay twitter messages such as colloquially language, novel words, and interjections into standard Malay language. This research will be used Language Model and N-grams model.

1. Introduction

Nowadays, there are so many social media or social networking for connecting people such as Facebook, Instagram, Whatsapp, WeChat, Twitter and so on. Dasselaar, (2016) mentioned that social media as an excellent mechanism for two-way communication[1]. Twitter is one of a free online social networking has above 500 million subscribers[2-3], it is also known as microblogging service that enables registered members to post a short note in 140 characters, beside that it also may contain photos, videos, links called as tweets, but to those unregistered Twitter only can read a tweet[3]. Thru twitter everyone either family, friends or officemates can communicate easily and also can stay connected through the exchange of quick and frequent messages[4].

Basically Twitter also is one of the top social networking sites in Malaysia, while normally registered user will use English or Malay language as media of communication. In fact, while use Malay language in twitter also had the same problem with casual written English language as per highlighted by Clark & Araki 2011 which is does not confirm the rules of spelling, grammar, punctuation, misspelled, use the abbreviation language and so on[5]. Meanwhile Samsudin et al. 2012 justify and identified have 13 patterns of common noisy terms in Malay Language[6]. Besides that, Hong, Convertino, & Chi, 2011 identified that Malay language also the higher leading language which used to communicate over the Twitter, Malay language is fourth leading language worldwide in twitter messages broadcasting[3-7]. Table 1 shows same sample of noisy text of Malay tweet.



Table 1.Sample noisy token Malay tweet

Original tweet: Akum Slh kew aq inGin cyunx duE OwAnk Yg BERBeZA RUPE?? Aq pown lyAl uNtOk coorUnx! Tutup akum
Malay normalize: Assalamualaikum Salah ke aku ingin sayang dua orang yang berbeza rupa? Aku pun layak untuk kau orang. Walaikumusalam
Translate: Hi It is wrong, if I fall in love with two difference people? I also eligible for you all. Bye

Based on table 1, can be justify there are several issues that makes the normalization of Malay tweet become a problematic and very difficult task to normalize all the texts. Based on this tweets conversation written extremely colloquially, novel words and interjections[8]. Besides that, a word written using a phonetic spelling (“idop” instead of hidup), and combined with other words into an acronym (“tutup akum” instead of walaikumussalam)[8]. Other than that, this tweet written with mixed uppercase and lowercase letters. Tweets need to be normalize and replace or remove as much as possible of noisy token or doing spell correction, then convert them to the standard Malay language.

2. Related works

The noisy text normalization methodologies can be written off as into five clusters. The SMT (statistical machine translation) paradigms firstly was familiarized by Aw, Zhang, Xiao and Su in 2006 and also become as the first paradigm’s in text normalization[9-10]. This paradigm is used to address normalization problems using statistical machine translation task. This approach visualize as a solution for paraphrase problem where a messages in the SMS language will be transformed to standard English language via a similar phrase-based statistical MT method [10]. In other word this paradigm use to normalize SMS text that translates a source language (UGC) to a target language (standard language)[4]. Time by time this technique has been upgraded and enhanced by further researches[11].

The second group of paradigms approaches is Spell-Checker metaphor. The Spell-Checker metaphor focusing on normalize SMS language, which is performed a correction on sequence of words and really concerns about out of vocabulary tokens from ‘noisy’ input token[12]. This noisy channel model metaphor was introduced by Choudhury, Saraf, Jain, Sarkar, and Basu in 2007, which this method was used a hidden Markov model (HMM)[12]. This method has been used to characterize the stochastic properties of the noisy channel. Further researches from other researchers has been refined and improved this method [13-14]. Besides that, this approach also have been modified by Cook and Stevenson (2009), the modification to design an unsupervised method using probabilistic models[3-15]. In addition, the spell checking approach have been merged with the SMT-like by Beaufort, Roekhaut, Cougnon, and Fairon (2010) for normalize French SMSs[16].

ASR metaphor represent as third group of paradigm approaches, this metaphor has been introduced by Kobus, Yvon, and Damnati 2008 . Kobus and others researcher introduced this metaphor to orthography of French SMS messages. This paradigm was combination of two metaphor which is an ASR-like system and SMT metaphor. Basically main purpose of this ASR metaphor to improve the accuracy of results normalize French SMS[17]. The fourth groups is the dictionary based normalization metaphor, which is very user friendly to used and able to come out fast result[4]. This approach requires the dictionary designed on the basis of the error and not proper English language used, commonly encountered in social media can be divided into several different categories, and in this regard, a multi-faceted approaches are the most effective way to address the problem[5]. It has been proven that using colloquial dictionary that can overcome some of the complex and sophisticated approach[4-5]. However, its performance is very depending on the size of the dictionary. Because of that, method to automatically compile a large data of dictionary has been introduced by Han, Cook, and Baldwin (2012)[18]. While identify the weaknesses in the dictionary approach, Oliva, Serrano, Del Castillo, and Igesias (2013) introduced a special Spanish phonetic dictionary, in which each entry

is formed by a coded consonant string, vowels strings, and their positions in the word, for normalizing Spanish SMS texts[4-19].

The fifth group is the Han and Baldwin who described lexical methods for normalize tweet message, this paradigm has been introduced in 2011. For this metaphor, Han and Baldwin implemented combination of four method for normalize the texts, which are; 1). Pre-processing; 2). Confusion set generation; 3). Ill-formed word detection and lastly 4). Candidate selection using a variety of metrics: lexical edit distance, phonemic edit distance, longest common subsequence (LCS), affix substring, language model, and dependency-based frequency features[20]. In 2011 also, Wei, et al enhanced HanBaldwin paradigm's to put an account of time sensitive query on twitter while doing normalization[21]. This encouraged us to plan a recovery methods which has three main phases: pre-processing, generation of candidates, and the selection of candidates[4].

3. Proposal Approach

3.1 Contextual Dictionary

Basically, most of people will using contextual cues to guessing the meaning of unfamiliar words[22]. Based on Knight, 1994, this research justified that normally dictionary will used into two major categories: vocabulary learning and reading comprehension[22]. In lines with the times, nowadays dictionary not only used for translating a language such as from Malay to English, but it also can translate colloquial language. Essentially, dictionary also will represented as database especially while using colloquial dictionary for normalize the texts.

Clark & Araki, 2011 was a pioneer in introduced dictionary metaphor to normalize the text. Their research tried to overcome the disambiguate word problem with the context-aware dictionary[5]. Meanwhile Basri et al. 2012 introduced dictionary normalization system which is contains eight modules in a pipeline for normalize Malay text[23]. After go through this research for every step of modules, can be identified that some of modules not suitable to be execute because will causes of data loosing and inaccurate normalization. Then at the same year Samsudin et al. 2012 ,develop a Malay noisy dictionary. They collecting almost 15000 data from various sources such as tweeter, online forum and Facebook after that manually normalize that words[6].

Basically based on research stated above, either from Clark, Basri or Samsudin and their researcher team, can be justified that this method or metaphor really depending on the size of dictionary database. Because of that, at the same year Han, et al. 2012 come out new method of research which are to extract the OOV and IV pairs based on distributional similarity from English Tweets then after that they select best pairs by string and contextual similarities of words[24]. The longer OOV words in sequence, this technique works more effective. Meanwhile based on Oliva, et al. 2013, this method proposed exercise of the lexical and semantic disambiguation using a combination of lexical resources such as SMS and Spanish dictionaries or the lexical database WordNet[19].

3.2 Decision Making via Language Model

Language model is one of the techniques in determining and compute the possibility of a sentence or sequences of word[24-27]. Originally language model was developed for overcome speech recognition problem[26]. After that this technique widely used in many areas also such as spelling correction, handwriting recognition, machine translation and optical character recognition[25]. A statistical language model be able to represent by the conditional probability of the next words based on sequence words. In the development of statistical model of natural language, the difficulty of this modelling problem is greatly reduced by taking advantage of word order, and the wordy fact that chronology of closer words are statistically dependant[28]. Therefore, n-gram models assemble tables of conditional prospect of the next word, for each one contexts large number[28].

Generally N-grams can be used in a multiplicity of different tasks. While developing the language model, normally n-gram is used to develop unigram model and it's not limited to use on that model only because it also can be used to develop bigram and trigram models. As example, Google and

Microsoft already urbanized n-gram models to the scale web, which are it can be used in a variation of tasks such as spelling correction, word-breaking and text identification. Besides that, from Gu & Berleant 2000 research, they identified that n-gram model also have been used as yearly as 1979 in so many task such as for language identification, spelling correction, document categorization, document comparison, robust handling of noisy (misspelled, OCR'ed etc.) texts, topic highlighting, document space visualization, spoken document retrieval, and other information retrieval related applications[29].

While doing the decision making via language model, feature selection must take an account also because feature selection is the process of selecting a subset of the conditions prevailing in the training set and only use this as a subset in text classification[30]. Feature selection has two main objectives. First, it makes training and use of more efficient classifiers to reduce the effective size of the vocabulary[30]. Second, the selection often improves the classification accuracy by eliminating noise characteristics[30]. Further explanation will elaborate at architecture section.

3.3 Architecture

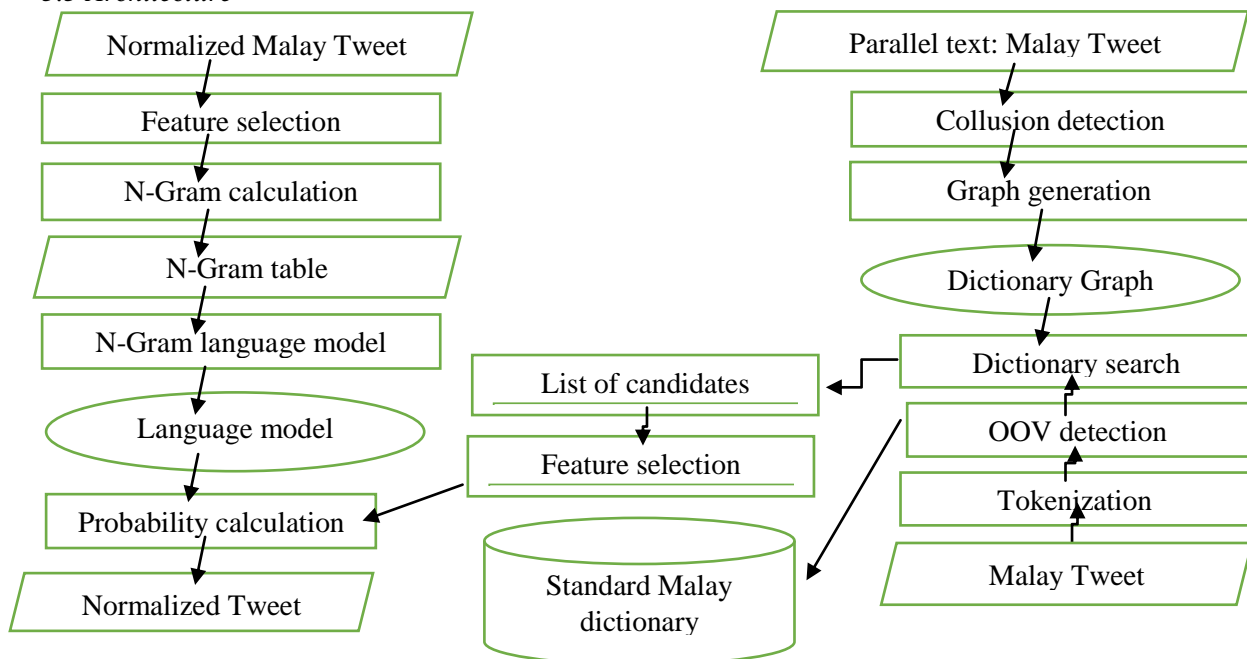


Figure 1. Proposed architecture of Statistical Dictionary Malay tweet normalization

Figure 1 represents proposed architecture of Hybrid Dictionary/Modelling Malay tweet normalization. The first important thing when doing normalization is to tokenize the text. As per suggested by Saloot, Idris & Mahmud 2014, in normalize Malay Twitter message are need to convert all capital letter in sentences into small letter, it because capital letter deos not have orthographic value[4]. This architecture also will implementing tokenization module has been introduced by Saloot, Idris & Mahmud 2014. Based on this architecture the second step will take an action is to descriminate OOV words and IV words. While detecting OOV words, we will checking the words by comparing with standard Malay dictionary database. In this OOV detection module will also implementing a set of XXX heuristic rules which has been introduced by Saloot, Idris & Aw 2014[3]. Besides that, C ++ vector data structure will be implementing to speed up the process and combined module responsible for generating candidate, modified lexical and phonemically[3]. This module is detected OOV sent word to the lexical components and applying a phoneme module to module lexical decision[3].

Actually when entering Malay tweet into this architecture, there are two jobs will process in parallel concurrently in order to get final result of normalized Malay tweet;1). Parallel text:Malay

Tweet; 2). Normalized Malay tweet. From the parallel text: Malay tweet proses flow sequences, first procedure is entering the Malay twitter message pair together with standard Malay language. Collusion detection method will used because it practice of two or more parties to work together where they have intentionally harm others[31]. While doing collusion detection, standard Malay language can be a benchmark for generating the graph, some algorithm will be create for graph generation[32]. From the study, we identified that Kronecker product of two matrices was the perfect instrument for achieve parallel text normalization goal[33]. After that dictionary graph will be created, the purpose of generate dictionary graph is for dictionary search, where it will work together with OOV detection, which is list of Malay words with the common proper name of cities and countries as example[3]. From the dictionary search, the list of candidates or result of the correct words will come out and again feature selection will be practices.

Meanwhile, at the same time in parallels this architecture also will be processing the second jobs which is normalization Malay tweet. Here, first feature selection will be implementing to making n-gram calculation. From the research Gu & Berleant 2000 justified because of uniqueness of strings of length n , n-gram can reduce the match word problem, also able to enhanced computational effectiveness compared with the word processing[34]. For the reason of the strings of n length, it can support until 5-gram even more and can be so challenging, that way hash array n-gram table need be create to store large amounts of texts[25-29]. Besides that, N-gram support partial matches if text contains words that are different but similar words because of the similarity between the words sequence having n-grams in common[29]. Creating N-gram tables was one of the step for making N-gram language modelling, because model whose assigning probabilities to sequence of words named as Language Model[28]. Actually what n-gram language model will do is to breaks up the sentence into smaller sequences of words and figures up the individual n-gram probabilities. The order of the words very important in Language Model and it can combine in order to create the LM by making weaker independence assumptions[34]. Zhu 2007, justified that the acclimatizing part $w_{i-n+1:i-1}$ is known as 'history', which $n - 1$ was represent previous words[34]. Basic formula for probability calculation n-gram language model. After make a probability calculation based on list of candidates from feature selection result, we can get final result of normalized texts.

4. Discussion

This research paper only presenting the conceptual level of architecture because it still work-in-progress research. The aim of this research is to develop a hybrid dictionary/modelling approach for Malay Tweet Normalization. Meanwhile this research have three main objectives: - 1). To propose new architecture based on statistical approach for normalizing Malay tweet. 2). To develop a prototype of hybrid dictionary model to normalize Malay tweet. 3). To evaluate the result/system (prototype) with BLUE evaluation technique and also validate the result of normalization Malay tweet. The architecture of this research will be contributed by my research on Malay Text Normalization.

5. Future work and Conclusion

This paper proposed a hybrid modelling/dictionary approach for Malay Tweet normalization. Based on purposed architecture/ modelling, this research will implementing machine learning which "gives computers the ability to learn without being explicitly programmed"[35]. This research target to handle all the problem or tasks while normalizing Malay Tweet likes supervised learning, unsupervised learning and reinforcement learning using hybrid dictionary and decision making via language model and n-grams model. For the future works, more detail and further experimental result elaborated in the next extended version of this paper.

References

- [1] Dasselaar, C., 2016. English Versus Native Language in Digital Diplomacy.
- [2] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the Demographics of Twitter Users.," in ICWSM, 2011.
- [3] Saloot, M.A., Idris, N. & Aw, A., 2014. Noisy Text Normalization Using an Enhanced Language Model. In Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition. Kuala Lumpur, Malaysia: SDIWC, pp. 111–122.
- [4] Saloot, M.A., Idris, N. & Mahmud, R., 2014. An architecture for Malay Tweet normalization. Information Processing and Management, 50(5), pp.621–633.
- [5] Clark, E. & Araki, K., 2011. Text normalization in social media : progress , problems and applications for a pre-processing system of casual English. , 27(Pacling), pp.2–11.
- [6] Samsudin, N. et al., 2012. Normalization of Common NoisyTerms in Malaysian Online Media. Proceedings of the Knowledge Management International Conference, (July), pp.515–520.
- [7] Hong, L., Convertino, G., & Chi, E. H. (2011). Language matters in Twitter: A large scale study. In L. A. Adamic, R. A. Baeza-Yates, & S. Counts (Eds.), ICWSM. The AAAI Press.
- [8] Kaufmann, M., 2010. Syntactic Normalization of Twitter Messages. , pp.1–7.
- [9] Aw, A., Zhang, M., Xiao, J., & Su, J. (2006). A phrase-based statistical model for SMS text normalization. In Proceedings of the COLING/ACL on main conference poster sessions (pp. 33–40). Stroudsburg, PA, USA: Association for Computational Linguistics.
- [10] Saloot, Idris, Shuib, R., (2015) Towards Tweets Normalization Using Maximum Entropy.
- [11] Lopez Ludeña, V., San Segundo, R., Montero, J. M., Barra Chicote, R., & Lorenzo, J. (2012). Architecture for text normalization using statistical machine translation techniques. In IberSPEECH 2012 (pp. 112–122). Madrid, Spain:
- [12] Kobus, C., Yvon, F. & Damnati, G., 2008. Normalizing SMS: Are Two Metaphors Better Than One? In Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1. COLING '08. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 441–448.
- [13] Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., & Basu, A. (2007). Investigation and modeling of the structure of texting language. International Journal of Document Analysis and Recognition (IJ DAR), 10(3), 157–174. <http://dx.doi.org/10.1007/s10032-007-0054-0>.
- [14] Liu, F., Liu, Y. & Weng, F., 2011. Why is "SXSX" trending ? Exploring Multiple Text Sources for Twitter Topic Summarization. Proceedings of the Workshop on Language in Social Media, (June), pp.66–75.
- [15] Xue, Z., Yin, D. & Davison, B., 2011. Normalizing Microtext. Analyzing Microtext, pp.74–79.
- [16] Cook, P. & Stevenson, S., 2009. An Unsupervised Model for Text Message Normalization. , (June), pp.71–78.
- [17] Beaufort, R. et al., 2010. A hybrid rule / model-based finite-state framework for normalizing SMS messages. , 1(July), pp.770–779.
- [18] Han, B., Cook, P. & Baldwin, T., 2012. Automatically constructing a normalisation dictionary for microblogs. EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference, (July), pp.421–432.
- [19] Oliva, J. et al., 2013. A SMS Normalization System Integrating Multiple Grammatical Resources. Natural Language Engineering, 19(1), pp.121–141.
- [20] Han, B. & Baldwin, T., 2011. Lexical Normalisation of Short Text Messages : Makn Sens a # twitter. , pp.368–378.
- [21] Z. Wei, L. Zhou, B. Li, K.-F. Wong, and W. Gao, "Exploring Tweets Normalization and Query Time Sensitivity for Twitter Search.," in TREC, 2011.
- [22] Knight, S., 1994. Dictionary Use While Reading : The Effects On Comprehension and Vocabulary Acquisition For Students Of Different Verbal Abilities. , (47).

- [23]Basri, S.B., Alfred, R. & On, C.K., 2012. Automatic spell checker for Malay blog. In Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on. pp. 506–510.
- [24]Collins, M., 2013. Language Modeling.
- [25]Goodman, J., 2001. A Bit of Progress in Language Modeling. Technical Report, p.73.
- [26]Watanabe, S. et al., 2011. Topic tracking language model for speech recognition. Computer Speech & Language, 25(2), pp.440–461.
- [27]Jurafsky, D., Language Modeling. , p.88.
- [28]Bengio, Y. et al., 2003. A Neural Probabilistic Language Model. The Journal of Machine Learning Research, 3, pp.1137–1155.
- [29]Gu, Z. & Berleant, D., 2000. Technical Report 10-00a, Oct. 2000, Software Research Lab, 3215 Coover Hall, Dept. of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011 USA. , pp.1–12.
- [30]Manning, C.D. & Raghavan, P., 2009. An Introduction to Information Retrieval A. C.-B. E. Salas, ed. Online.
- [31]Mazrooei, P., 2013. Collusion detection in sequential games. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, MR90652, p.88
- [32]Cordeiro, D. et al., 2010. Random graph generation for scheduling simulations. Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques, p.60.
- [33]Leskovec, J., Chakrabarti, D. & Kleinberg, J., Realistic , Mathematically Tractable Graph Generation and Evolution , Using Kronecker.
- [34]Zhu, X., 2007. CS838-1 Advanced NLP : Language Modeling The Need for Finding Likely Sentences. Language, (3), pp.1–7.
- [35]"Machine Learning: What it is and why it matters". *www.sas.com*. Retrieved 2016-12-29.