Microtext Normalization for Chatbots

Ranjan Satapathy[1,2], Erik Cambria[2], and Nadia Magnenat Thalmann[1]

[1] Institute for Media Innovation, Nanyang Technological University, Singapore
{satapathy.ranjan, nadia}@ntu.edu.sg
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore
cambria@ntu.edu.sg

Abstract. This study aims at enhancing the human-computer interaction by incorporating a microtext lexicon and decreasing the response time by adding a binary classifier. Microtext lexicon and binary classifier together constitute the microtext module. The work leverages on the fact that humans tend to write in different unconstrained ways. Such unconstrained ways of communication comes under the umbrella of microtext analysis. Here microtext normalization technique is incorporated into a chatbot. The results show an improvement in the chatbot's understanding to any form of unconstrained languages. The Bilingual Evaluation Understudy score is used to evaluate the efficiency before and after normalization. Results show that the microtext module promises to increase both unconstrained text (SMS) and social media language (tweets) understanding.

Keywords: Microtext Normalization · Chatbot · Dialogue System.

1 Introduction

Building a dialogue system which understands human language is not an easy task as the humans interact socially in enormous different ways. Communicating using unconstrained natural language is an intuitive and flexible way for humans to interact. Understanding this kind of linguistic input is challenging for machines because of the diversity found in words and phrases used over different social media platforms. In order to interact with and understand humans, machines need to understand the different unconstrained ways people write. The popularization of mobile phones and social networks, is evident from the frequency of tweets which has reached an astonishing figure of more than 8,000 tweets produced per second[3]. There are many abbreviations and non-standard words used in SMSs and tweets [14]. These type of communications are usually performed in real time and over platforms which impose limits on the length of the messages, as in the case of Twitter and the traditional SMS system. Due to these constraints, the writing format of these messages clearly differs from normal standards. Features such as word shortenings, contractions and abbreviations are commonly used both to gain writing speed and circumvent the length limitations.

In recent years, the rise and expansion of social media has enabled users to share their views and interests in an impromptu manner. For example, they write terms or sentences such as "c u 2morrow" (see you tomorrow), "tgif" (thank God it's Friday) and

[3] http://www.internetlivestats.com/one-second/

“abt” (about) which may not be found in standard English but are widely seen in SMS, tweets, Facebook posts, blogs, discussion forums and chat logs. These unconstrained ways of writing text is called microtext. Microtext became one of the most widespread communication forms among users due to its casual writing style and colloquial tone [17].

The rise of social media usage has also led to the unconstrained generation of sentences in speech such as “wassup” (what is up), “howz” (how is) and interjections like ahem, aw, etc. which has emotions attached to them. Given that most data today is mined from the web, microtext analysis is key for many natural language processing (NLP) and data mining tasks, as most text classifiers are trained in plain English. In the context of sentiment analysis, microtext normalization is a necessary step for pre-processing text before polarity detection is performed [4].

The challenge arises when systems try to automatically rectify and replace them with the standard words [18,23]. Microtext normalization could be thought of as a simple find-and-replace pre-processing [12] step. For instance, a sampling of Twitter studied in [18] found over 4 million OOV words where new spellings were created constantly, both voluntarily and accidentally.

- **Input Text** : Wassup Nadine[a]
  - **chatbot’s actual answer**: I could not find an answer to that.
  - **Expected chatbot’s answer** : I’m doing good. How about you
- **Input Text** : Howz you doing
  - **chatbot’s actual answer** : I could not find an answer to that.
  - **Expected chatbot’s answer** : I am doing good. How about you?
- **Input Text** : Talk to you later
  - **chatbot’s actual answer** : Talk to you later
  - **Expected chatbot’s answer** : Talk to you later

[a] Nadine is the name of chatbot used for conversation

The proposed work is a step towards curbing the gap between the humans and chatbot by leveraging on a microtext lexicon to transform out-of-vocabulary (OOV) words to their in-vocabulary (IV) or human readable counterparts. The rest of the paper is as follows: Section 2 explains the related work, Section 3 explains the proposed framework, Section 3.1 explains the Datasets used, Section 4 explains the results and discussions and finally the Section 5 explains the conclusion and future work.

2 Related Work

Opinions and its associated concepts such as sentiments, emotions, attitudes, and evaluations are the center of study of sentiment analysis. This section discusses through the related work in microtext normalization and dialogue systems.

2.1 Microtext Analysis

Microtext has become ubiquitous in today's communication. This is partly a consequence of Zipf's law, or principle of least effort (for which people tend to minimize energy cost at both individual and collective levels when communicating with one another), and it poses new challenges for NLP tools which are usually designed for well-written text [10]. Normalization is the task of transforming unconventional words/sentences to their respective standard counterpart.

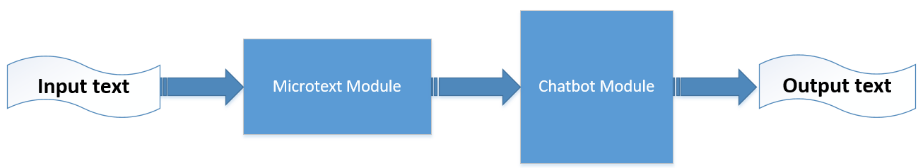


Fig. 1: Proposed framework for Chatbot

In [16], authors present a novel unsupervised method to translate Chinese abbreviations. It automatically extracts the relation between a full-form phrase and its abbreviation from monolingual corpora, and induces translation entries for the abbreviation by using its full-form as a bridge. [9] uses a classifier to detect OOV words, and generates correction candidates based on morphophonemic similarity. The types and features of microtext are reliant on the nature of the technological support that makes them possible. This means that microtext will vary as new communication technologies emerge. In our related work, we categorized normalization into three well-known NLP tasks, namely: spelling correction, SMT, and automatic speech recognition (ASR).

Spelling Correction Correction is executed on a word-per-word basis seen as a spelling checking task. This model gained extensive attention in the past and a diversity of correction practices have been endorsed by [6,3,15,21,27]. Instead, [26] and [7] proposed a categorization of abbreviation, stylistic variation, prefix-clipping, which was then used to estimate their probability of occurrence. Thus far, the spelling corrector became widely popular in the context of SMS, where [5] advanced the hidden Markov model whose topology takes into account both "graphemic" variants (e.g., typos, omissions of repeated letters, etc.) and "phonemic" variants (e.g., spellings that resemble the word's pronunciation).

Statistical Machine Translation Statistical Machine Translation (SMT) outlooks microtext as a foreigner language that has to be translated to plain English, meaning that normalization is done through a SMT task. When compared to the previous task, this method appears to be rather straightforward and better since it has the possibility to model (context-dependent) one-to-many relationships which were out-of-reach previously [13]. Some examples of works include [1,11,22]. However, the SMT still overlooks some features of the task, particularly the fact that lexical creativity verified in social media messages is barely captured in a stationary sentence board.

Automatic Speech Recognition ASR considers that microtext tends to be a closer approximation of the word's phonemic representation rather than its standard spelling. As follows, the key of microtext normalization becomes very similar to speech recognition which consists of decoding a word sequence in a (weighted) phonetic framework. [13] proposed to handle normalization based on the observation that text messages present a lot of phonetic spellings, while more recently [12] proposed an algorithm to determine the probable pronunciation of English words based on their spelling. Although the computation of a phonemic representation of the message is extremely valuable, it does not solve entirely all the microtext normalization challenges (e.g., acronyms and misspellings do not resemble their respective IV words' phonemic representation). Authors in [2] have merged the advantages of SMT and the spelling corrector model.

2.2 Dialogue System

Authors in [30] built an open-domain end-to-end human-computer conversational agent to integrate a large commonsense knowledge base into end-to-end conversational models. [25] investigated the limitations of building a Generative Hierarchical Neural Network Models based dialogue system and show how it outperforms state-of-the-art neural language models. Emotion detection in conversations [19] is a necessary step for a number of applications, including opinion mining over chat history, social media threads, debates, argumentation mining, understanding consumer feedback in live conversations,etc. Currently systems do not treat the parties in the conversation individually by adapting to the speaker of each utterance. There are social media based chatbot [29,8] which do not take microtexts into account. So, our main motive is to include the microtexts to train the system, so that the chatbot learns the intrinsic linguistic patterns and generate a response accordingly.

3 Proposed framework for chatbot

Composite nature of the NLP problem is addressed by the suitcase model [4]. In this regard, microtext module is the first step. The syntactics layer aims at preprocessing text so that informal text is reduced to human readable format (any language), inflected forms of verbs and nouns are normalized, and basic sentence structure is made explicit. Though, we could always build a rule based system to handle such events, but social media language is dynamic. It incorporates new short forms rapidly.In order to update the lexicon, we crawled popular acronyms from NetLingo[4], MakeUseOf[5], Slangs[6], and Internet Slang[7].

Table 1: Sample Lexicon incorporated in Nadine's system

OOV word	Class	Polarity	IV word
a3	OTHER	NEUTRAL	anytime, any place, anywhere
ru/18	OTHER	NEUTRAL	are you over 18?
AAF	ACR	POSITIVE	As A Friend
bestie	OTHER	POSITIVE	best friend
ne1	PHON	NEUTRAL	anyone
urz	PHONETIC	NEUTRAL	yours
b3	OTHER	NEGATIVE	blah, blah, blah
aight	CLP	NEUTRAL	all right

Microtext is divided into 5 classes based on the features it possess. The classes are as follows:

1. Clipping
2. Phonetic
3. Acronym
4. Hybrid
5. Others

The proposed model incorporates microtext understanding in the chatbot. It helps the chatbot to understand the unconstrained languages as shown in Table 2. The framework shown in Figure 2 has a binary classifier which classifies a text into OOV or IV, based on the learned features. The classifier employs a n-gram model with several machine learning techniques as shown in Table 3a and Table 3b.

Table 1 shows the sample lexicon which helps social robot's NLP module understand the social media language. The proposed framework is shown in Figure 1. The text is passed through microtext module for normalization and then passed on to Nadine's NLP module.

Microtext	Meaning	Polarity
aah	Fright	NEGATIVE
aha	Understanding, triumph (can also be used as "ahh")	POSITIVE
duh	Expresses annoyance over something stupid or obvious	NEGATIVE
haha	Regular laughter	POSITIVE
wow	Impressed, astonished	POSITIVE

Table 2: Examples of unconstrained language with emotions associated with it

[4] Reproduced by Permission©1995-2018 NetLingo®The Internet Dictionary at http://www.netlingo.com
[5] http://makeuseof.com/tag/30-trendy-internet-acronyms
[6] http://acronymsandslang.com/
[7] http://internetslang.com/

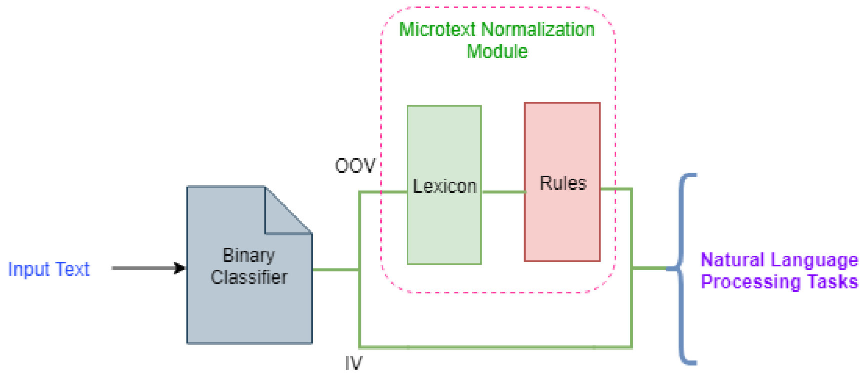


Fig. 2: Proposed framework

3.1 Datasets

This section discusses the datasets used in the evaluation of proposed framework. The unconstrained style of speech comes under the umbrella of microtexts. The two classes in both the datasets are equally distributed.Table 3b and Table 3a shows the accuracy of machine learning algorithms on different datasets.

Table 3: Evaluation results on different datasets

(a) 10-fold Accuracy on NUS SMS dataset

Classifier	**10-fold (%)**
NuSVC	85.14
Linear SVC	***92.95***
Original Naive Bayes	89.62
Multinomial Naive Bayes	89.92
Bernoulli Naive Bayes	89.24
Logistic Regression	91.05
SGDC	91.42

(b) 10-fold Accuracy on Normalized tweets dataset

Classifier	**10-fold (%)**
NuSVC	84.2
Linear SVC	***87.4***
Original Naive Bayes	83.5
Multinomial Naive Bayes	81.8
Bernoulli Naive Bayes	82.7
Logistic Regression	84.9
SGDC	83.4

NUS SMS Corpus This corpus (Table 4) has been created from the NUS English SMS corpus[8], the authors [28] randomly selected 2,000 messages. The messages were first normalized into standard English and then translated into standard Chinese. For our

[8] http://github.com/kite1988/nus-sms-corpus

evaluation purposes, we only used the actual messages and their normalized English version (leaving out their Chinese counterparts).

Social media texts	Expanded forms
I'll meet u b4 lec then...	I will meet you before the lecture then.
Where r u	Where are you
Hey are we going out tmr	Hey are we going out tomorrow
So u stayin in d hostel ?	So you are staying in the hostel ?
R u going to b done anytime soon ?	Are you going to be done anytime soon ?

Table 4: Sample real time tweets/SMS

Normalized Tweet Dataset Authors in [24], built a lexicon which consists of real time tweets and their IV counterparts. The dataset is available on request.

4 Results and Discussion

The Bilingual Evaluation Understudy score (BLEU) score is used to evaluate the sentences' similarity. The Sentence BLEU[9] is used to score the similarity between normalized sentences output from the proposed framework and human annotated sentences.

4.1 Dataset Collection and Annotation

The dataset in [24] was available on request. The dataset consists of tweets crawled from Twitter streaming API[10]. The data was preprocessed using following rules:

1. removal of usernames (starting with),
2. urls (eg., https://www.Twitter.com),
3. Removal of punctuation marks,

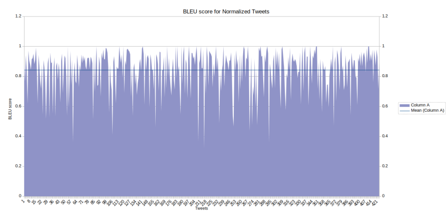
4.2 Time Complexity

The results in Table 3a and Table 3b shows different algorithms applied on both the datasets. The models are trained on the unigram features as microtexts work at the word-level [24]. The result shows **Linear SVC** to be the best binary classifier for both the datasets. The binary classifier reduces the time complexity and makes the overall framework run faster. The framework ran on Python 3.5 on Ubuntu operating system with 64 GB RAM and 30 GB 1080 T_i Nvidia Graphics. It took 11.4 seconds to run without the binary classifier and only 8.8 seconds with the binary classifier. The binary classifier works as a filter, which reduces the overall execution time of the framework.

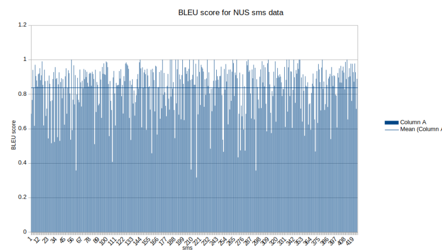
[9] https://www.nltk.org/_modules/nltk/translate/bleu_score.html
[10] https://developer.twitter.com/en/docs

4.3 BLEU score

BLEU score [20] is employed as an evaluation task. It is used to evaluate the quality of text which has been machine-translated from one natural language to another. It's strength is that it correlates highly with human judgements by averaging out individual sentence judgment errors. Figure 3a and Figure 3b shows the BLEU score for the normalized Tweet and NUS SMS data respectively. The results show **Mean BLEU score** of more than **0.8** is achieved for both the dataset. The model's output is compared against the human annotated text as provided in the datasets.



(a) BLEU score for the normalized Tweets data



(b) BLEU score for the NUS SMS data

Fig. 3: Evaluation of datasets based on BLEU score

5 Conclusion and Future Work

The proposed framework consists of a binary classifier which classifies a given sentence into either microtext or non-microtext. Binary classifier takes syntactic features to determine a class label. Linear SVC gives an accuracy of 87.4% on Normalized Tweet Dataset and 92.95% on NUS SMS data. The addition of binary classifier also improves the overall execution time of the task. The detected microtexts are then passed through the lexicon. Lexicon transforms the out-of-vocabulary texts to their in-vocabulary counterparts. BLEU score was taken as an evaluation metric and shows a mean of more than 0.8 in both the datasets. Future work will focus on experimenting whether lexicons could be replaced by a more cognitive approach which is a phonetic system (e.g., International Phonetic Alphabet). It will improve the generalization of the proposed rules based on more cognitive qualities of speech such as phones, phonemes, intonation and separation of words and syllables.

Acknowledgment

This research is supported by the BeingTogether Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at

Chapel Hill. The BeingTogether Centre is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

1. Aw, A., Zhang, M., Xiao, J., Su, J.: A phrase-based statistical model for SMS text normalization. In: 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. pp. 33–40 (2006)
2. Beaufort, R., Roekhaut, S., Cougnon, L.A.l., Fairon, C.d.: A hybrid rule/model-based finite-state framework for normalizing SMS messages. In: ACL. pp. 770–779. Association for Computational Linguistics (2010)
3. Brill, E., Moore, R.C.: An improved error model for noisy channel spelling correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. pp. 286–293 (2000)
4. Cambria, E., Poria, S., Gelbukh, A., Thelwall, M.: Sentiment analysis is a big suitcase. IEEE Intelligent Systems **32**(6), 74–80 (2017)
5. Choudhury, M., Saraf, R., Jain, V., Sarkar, S., Basu, A.: Investigation and modeling of the structure of texting language. International Journal of Document Analysis and Recognition **10**(3-4), 157–174 (2007)
6. Church, K.W., Gale, W.A.: Probability scoring for spelling correction. Statistics and Computing **1**(2), 93–103 (1991)
7. Cook, P., Stevenson, S.: An unsupervised model for text message normalization. In: Proceedings of the workshop on computational approaches to linguistic creativity. pp. 71–78 (2009)
8. Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., Zhou, M.: Superagent: a customer service chatbot for e-commerce websites. Proceedings of ACL 2017, System Demonstrations pp. 97–102 (2017)
9. Han, B., Baldwin, T.: Lexical normalisation of short text messages: Makn sens a# twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 368–378 (2011)
10. Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media. pp. 216–225 (2014)
11. Kaufmann, M., Kalita, J.: Syntactic normalization of twitter messages. In: International conference on natural language processing, Kharagpur, India (2010)
12. Khoury, R.: Microtext Normalization using Probably-Phonetically-Similar Word Discovery. In: Wireless and Mobile Computing, Networking and Communications (WiMob), 2015 IEEE 11th International Conference on. pp. 392–399 (2015)
13. Kobus, C., Yvon, F., Damnati, G.: Normalizing SMS: are two metaphors better than one? In: Proceedings of the 22nd International Conference on Computational Linguistics. vol. 1, pp. 441–448. Association for Computational Linguistics (2008)
14. Li, C., Liu, Y.: Normalization of text messages using character-and phone-based machine translation approaches. In: Thirteenth Annual Conference of the International Speech Communication Association. pp. 2330 – 2333 (2012)
15. Li, M., Zhang, Y., Zhu, M., Zhou, M.: Exploring Distributional Similarity Based Models for Query Spelling Correction. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. pp. 1025–1032. ACL-44, Association for Computational Linguistics (2006)

16. Li, Z., Yarowsky, D.: Unsupervised translation induction for chinese abbreviations using monolingual corpora. In: Proceedings of ACL-08: HLT. pp. 425–433. Association for Computational Linguistics (2008)
17. Liu, F., Weng, F., Jiang, X.: A broad-coverage normalization system for social media language. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 1035–1044. Association for Computational Linguistics (2012)
18. Liu, F., Weng, F., Wang, B., Liu, Y.: Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies **2**, 71–76 (2011)
19. Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., Cambria, E.: DialogueRNN: An attentive RNN for emotion detection in conversations. In: Thirty-third AAAI Conference on Artificial Intelligence (2019)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
21. Pennell, D., Liu, Y.: A character-level machine translation approach for normalization of sms abbreviations. In: Proceedings of 5th International Joint Conference on Natural Language Processing. pp. 974–982 (2011)
22. Pennell, D.L., Liu, Y.: Normalization of informal text. Computer Speech & Language **28**(1), 256–277 (2014)
23. Petrović, S., Osborne, M., Lavrenko, V.: The Edinburgh Twitter corpus. In: Proceedings of the NAACL HLT Workshop on Computational Linguistics in a World of Social Media. pp. 25–26 (2010)
24. Satapathy, R., Guerreiro, C., Chaturvedi, I., Cambria, E.: Phonetic-based microtext normalization for twitter sentiment analysis. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 407–413. IEEE (2017)
25. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. pp. 3776–3783. AAAI Press (2016)
26. Sproat, R., Black, A.W., Chen, S., Kumar, S., Ostendorf, M., Richards, C.: Normalization of non-standard words. Computer speech & language **15**(3), 287–333 (2001)
27. Toutanova, K., Moore, R.C.: Pronunciation Modeling for Improved Spelling Correction. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 144–151. Association for Computational Linguistics (2002)
28. Wang, P., Ng, H.T.: A beam-search decoder for normalization of social media text with application to machine translation. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 471–481 (2013)
29. Xu, A., Liu, Z., Guo, Y., Sinha, V., Akkiraju, R.: A new chatbot for customer service on social media. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 3506–3510. ACM (2017)
30. Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., Huang, M.: Augmenting end-to-end dialogue systems with commonsense knowledge. In: Thirty-Second AAAI Conference on Artificial Intelligence. pp. 4970–4977 (2018)