SMS Text Normalization Using Hybrid Approach

Sakshi Goyal[1] , Er.charandeep Singh Bedi[2]

[1]M.Tech Final Year Student
Deptt. of Computer Science & Engg. BFCET, Bathinda,India.
[2]Assistant Professor
Deptt. of Computer Science & Engg. BFCET, Bathinda,India.

***Abstract*:** Text normalization is a task of generating plain text from an un normalized text. Mobile technology has contributed to the evolution of several media of communication such as chats, emails and short message service (SMS) text. This has significantly influenced the traditional standard way of expressing views from letter writing to a high-tech form of expression known as texting language. In this paper we present a review on various techniques to used translate a un normalized text into its equivalent plain text.

Keywords: **Text Normalization, Statistical Machine Translation Approach, Nearest word locater, Dictionary lookup.**

I.INTRODUCTION

Text normalization is a process of translating SMS text into plain English. There has been a rapid increase in social text in the last few years, including the mobile phone text messages (SMS), comments from the social media websites such as Facebook and Twitter, and real-time communication platforms like MSN and Gtalk [2].Knowledge from this data. Unfortunately, traditional NLP tools sometimes perform poorly when processing this kind of text. One of reasons is that social text is very informal, and contains many misspelled words, abbreviations and many other non-standard tokens. Short Messaging Service (SMS) texts behave quite differently from normal written texts and have some very special phenomena. To translate SMS texts, traditional approaches model such irregularities directly in Machine Translation (MT). However, such approaches suffer from customization problem as tremendous effort is required to adapt the language model of the existing translation system to handle SMS text style. We offer an alternative approach to resolve such irregularities by normalizing SMS texts before MT. In this thesis work, we view the task of SMS normalization as a translation problem from the SMS language to the English language and we propose statistical MT model for the task. The problem of text normalization can be explained with the help of an example. Consider a SMS text "**shd we go 2 yr house den?**" this SMS text can be normalized in the plain English as **"Should we go to your house then ?".**

II. LITERATURE SURVEY

Karthik Raghunathan, Stefan Krawczyk[1]:. This paper explores two approaches to SMS text normalization. First,a dictionary substitution approach used by most websites that provide such a service, and then modify it with extension. It ends the discussion about the shortcomings of the system and possible improvements in the future to make it better**.**

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, Cédrick Fairon, July 2010 [2]:This paper presents a method that shares similarities with both spell checking and machine translation approaches. The normalization part of the system is entirely based on models trained from a corpus. The principle of this method of evaluation is to split the initial corpus into 10 subsets of equal size. The system is then trained 10 times, each time leaving out one of the subsets from the training corpus, but using only this omitted subset as test corpus .The language model of the evaluation is a 3-gram. System did not try a 4-gram. Overall accuracy of the system is comes out to be 76.23%.

Chen Li Yang Liu, Improving Text Normalization Using Character-blocks based Models and System Combination[3]:In this paper, author propose an approach to segment words into blocks of characters according to their phonetic symbols, and apply MT and sequence labeling models on such block-level. Author also proposes to combine these methods, as well as with other existing methods, in order to leverage their different strengths. The proposed system shows an accuracy of 74.6%.

DEANA L. PENNELL, B.S.,M.S, NORMALIZATION OF INFORMAL TEXT FOR TEXT-TO-SPEECH 2010 IEEE

[4]: A large amount of information is found in noisy contexts as texting and chat lingo have become increasingly prolific in the past decade. The increase in performance is seen even when tested solely on deletion abbreviations, meaning that the MT system is able to utilize some feature in the text that is not covered by those used in this work.

III. METHODOLOGY

To translate the SMS text into its equivalent plain text following approaches can be used.

A. Statistical machine translation approach:

Statistical machine translation is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge. During translation, the collected statistical information is used to find the best translation for the input sentences, and this translation step is called the decoding process. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model. Statistical MT model take the view that every sentence in the target language is a translation of the source language sentence with some probability. The best translation, of course, is the sentence that has the highest probability. This technique will work Statistical Machine Translation (SMT) approach works in two phases (1) Training phase (2) Translation phase. Training phase is used to extract seven tables like Bi-gram, Tri-gram, Four-gram, Five-gram, Six-gram, Seven-gram with the help of uni-gram table and parallel corpus provided for training. In Translation phase actual translation is done for the end user with the help of these seven tables extracted during training phase.

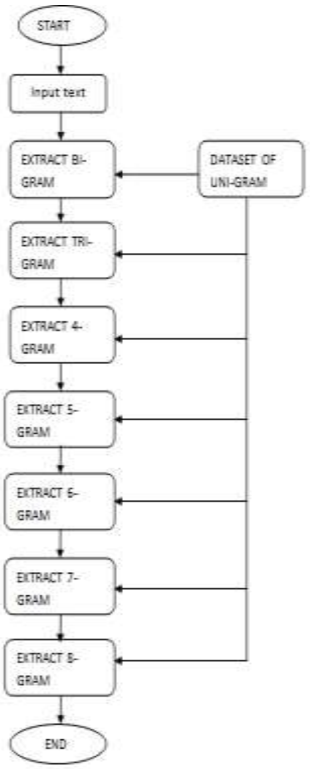


Fig 1:Flowchart of Training phase

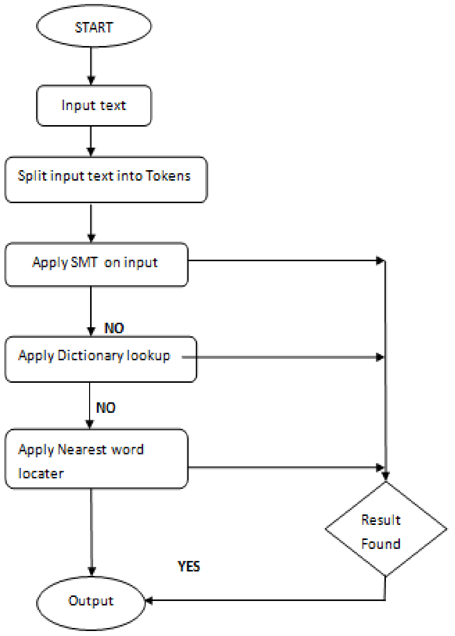


Fig 2:Flowchart of Translation phase

B. Dictionary Lookup Technique: This approach is mainly used to check whether the particular token is correct or not by comparing the token with the dictionary values. It is assumed that the word which is being checked is correct if it is available in the dictionary. To create dictionary for various English words, various resources like English text books, online English websites are being used. The accuracy of the system is highly depends upon this phase. If the required word is correct but not in the dictionary then it will give wrong output

C. Nearest Word Locater Technique: This technique will work if dictionary lookup approach becomes unable to generate the accurate word. This technique is used to find the nearest possible word from the dictionary to obtain the result. With the help of this technique various suggestions are generated with respect to the token which is being checked in the ascending order of their distances. In this approach, the word distance means the minimum number of operations required to equate the wrong word with the word in dictionary.

III. RESULTS AND DISCUSSIONS

Proposed system is tested on 200 different short messages on SMT by using Uni-gram Table with the help of the database available from karthik Raghunathan successfully. System is also manually tested on 400 sentences out of which 380 are converted successfully and hence proposed system shows the accuracy of 83% which shows that the results are quite good.

Table I: Dictionary look up approach

Abb	**Eng**
Dere	There
Wat	What
Gng	Going
Wry	Worry
Tnsn	Tension
Uni	University
Bti	Bathinda
Chd	Chandigarh

The following table shows the results generated by the system:

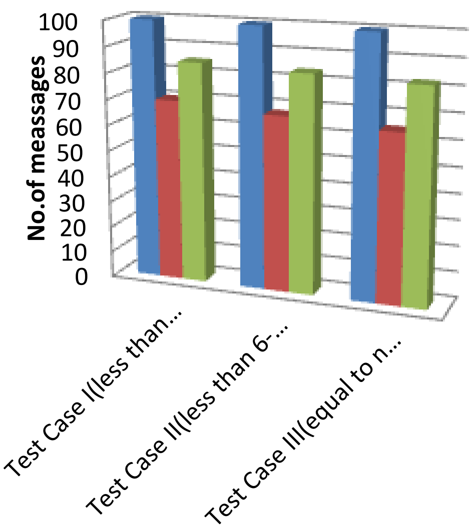
Table II: RESULTS of SMT

Input Sentence	**Output Generated by System**
I m here	I am here
R u fi9	Are you fine
I m gng to uni	I am going to university
gud 4 nthing	good for nothing
hlo I m gng to mt my frend	hello I am going to meet my friend
b dere	be there
ASAP	As soon as possible

The following table shows the statistics of proposed system

Table III: Results of system

Parameter	Value
English Dictionary Words	1,50000
Abbreviations	2000
Sample short message	200
Tested Inputs	400
Accurately Converted	378
Overall System Accuracy	83%



	Test Case I(less than equal to 3-grams)	Test Case II(less than 6-grams)	Test Case III(equal to n above 6-grams)
🟦 Precison	100	100	100
🟥 Recall	70	67.5	65
🟩 F-Measure	85	83.75	82.5

Table IV: Results of Test-Cases

IV. Conclusion

Proposed system can be used to translate the short message into its equivalent plain English text. In proposed system hybrid approach is used to translate the short message into its equivalent plain English text. We have defined normalization in this System. System presents various techniques like dictionary lookup, Nearest word locator and SMT of translation of un normalized text into plain English text. System have used up-to 8-grams of input. Maximum accuracy of the existing systems comes out to be 83.75% according to our database. Results of the proposed system shows some improvement over the results of the existing system.

FUTURE SCOPE

In Future work, SMT approach can be further extended to improve the accuracy of the overall system. It can be used to translate multi languages. Large corpus can be developed to increase the accuracy. Parallel corpus of SMT sentences can also be increased further to achieve more accurate results

ACKNOWLEDGEMENT

Proposed system is tested on 200 different short messages on SMT by using unigram Table with the help of the database available from karthik Raghunathan[1] successfully.

REFERENCES

[1]Stefan Krawczyk Karthik Raghunathan. Investigating sms text normalization using statistical machine translation.Stanford University,Stanford, CA.

[2]RichardBeaufort, Sophie Roekhaut, Louise-Amélie Cougnon, Cédrick Fairon. A hybrid rule/model-based finite-state framework for normalizing SMS messages.

[3]ChenLi Yang Liu,Improving Text Normalization Using Character-blocks based Models and SystemCombination..

[4]Noam Chomsky and Morris Halle. 1968. The sound pattern of English. Harper and Row, New York, NY.

[5]AiTi Aw, Min Zhang, Juan Xiao, Jian Su A Phrase-based Statistical Model for SMS Text Normalization Institute of Infocomm Research 21 Heng Mui Keng Terracegapore 119613.

[6]Ademola O. Adesina, Kehinde K. Agbele, Nureni A. Azeez, Ademola P. Abidoye , A Query-Based SMS Translation in Information Access System .

[7]Automatic normalization of short texts by combining statistical and rule-based techniques.Architecture for Text Normalization using Statistical Machine Translation techniques.

[8]w, AiTi and Zhang, Min and Xiao, Juan and Su, Jian,"A phrase-based statistical model for SMS text normalization", Proceedings of the COLING/ACL on Main conference poster sessions,2006, pages 33–40, Sydney, Australia.

[9]Catherine Kobus, François Yvon, and Géraldine Damnati. Normalizingsms: are two metaphors better than one? In COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics, pages 441–448, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[10]A Machine-Translation Method for Normalization of SMS Darnes Vilari˜no, David Pinto, Beatriz Beltr´an, Saul Le´on, Esteban Castillo, and Mireya Tovar

[11]Tim Schlepped, Chenfei Zhu, Jan Gebhardt, Tanja Schultz Text Normalization based on Statistical Machine Translation and Internet User Support Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany.

[12]Congle Zhang Adaptive Parser-Centric Text Normalization Dept of Computer Science and Engineering University of Washington, Seattle, WA 98195, USA.

[13] Choudhury, Monojit, Rahul Saraf, Vijit Jain, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language.