A REVIEW PAPER ON SMS TEXT TO PLAIN ENGLISH TRANSLATION(Text Normalization)

*MEENAKSHI SHARMA
M.Tech Final Year Student (Department of Computer Science)
Giani Zail Singh PTU Campus, Bathinda
Bathinda (Punjab)
ermeenakshi89@gmail.com

**DR. PARAMJEET SINGH
Assistant Professor (Department of Computer Science)
Giani Zail Singh PTU Campus, Bathinda
Bathinda (Punjab)
param2009@yahoo.com

***DR. SHAVETA
Assistant Professor (Department of Computer Science)
Giani Zail Singh PTU Campus, Bathinda
Bathinda (Punjab)
Garg.shavy@yahoo.com

ABSTRACT

Mobile technology as well as social networking technology plays an important role in communication across internet. A large amount of information is found in noisy contexts as texting and chat lingo have become increasingly considerably in the past decade. This noisy information needs to be normalized into the standard text so that it can be used by the various other tools such as text-to-speech programs. This paper presents a review on Short message Service (SMS) text normalization into plain English text. Term normalization means to translate the SMS text into the plain English text using various techniques like Rule based approach and Statistical machine translation. This is research area of Natural Language Processing (NLP).

Keywords: Rule based Approach; Statistical Machine Translation; Text Normalization; Translation

I. INTRODUCTION

Today's world is an era of information technology in which every person wants to use such internet tools to be connected with one another. In this path mobile technology and other chatting and communication activities plays an important role. Users of today's world use these technologies to pass their messages , view points and sometimes some important information with the help of these technologies. While typing a message or some text user often uses some shorthand expressions or uses such language which contains a lot of noise in their important text formation? Users can also store their data and knowledge in this format. To take benefit from this text various natural language processing techniques need to be adapted to work accurately on this unconventional data. This paper describes various approaches for the normalization of informal text, such as that found in emails, social networking messages, chat rooms, and SMS messages. For example a text message can be written as follows:

"**I wanna be der**"

Can be translated in the plain English as

"**I want to be there**".

For normalization or translation purpose various techniques can be used such as Rule based approach, Statistical Machine translation Approach along with the parallel corpus. Some websites are also providing the SMS text to plain English translation approaches but technology behind these websites is only dictionary lookup approach which means they just replace the words from the input text by their equivalent target sentence by finding from the dictionary. They are not using any language model for translation purpose. The English signal is sent across a noisy channel, as an SMS, which we then try to recover using a language and translation model along with the help of Natural Language Processing (NLP) techniques.

II. LITERATURE SURVEY

Deana L. Pennell and Yang Liu, Normalization Of Text Messages For Text-To-Speech

This paper describes a normalization system for text messages to al- low them to be read by a TTS engine. To address the large number of texting abbreviations, author use a statistical classifier to learn when to delete a character. The features we use are based on character context, function, and position in the word and containing syllable. To ensure that our system is robust to different abbreviations for a word, system generate multiple abbreviation hypotheses for each word based on the classifier's prediction. System then reverse the mappings to enable prediction of English words from the abbreviations. Results show that this approach is feasible and warrants further exploration. Author evaluate classifier accuracy by performing 10-fold cross validation on the training data. Always choosing the positive class System yields a baseline accuracy of 74.7%. [1]

Richard Beaufort , Sophie Roekhaut , Louise-Amélie Cougnon, Cédrick Fairon, A hybrid rule/model-based finite-state framework for normalizing SMS messages

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. Evaluated in French by 10-fold-cross validation, the system achieves a 9.3% Word Error Rate and a 0.83 BLEU score. The evaluation was performed on the corpus of 30,000 French SMS presented in Section 4.2, by ten-fold cross-validation (Kohavi, 1995). 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%. [2]

ChenLi Yang Liu,Improving Text Normalization Using Character-blocks based Models and System Combination

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. Authors also propose 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%. [3]

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

In this paper author investigated building a mobile information access system based on SMS queries. The difficulties with SMS communication were explored in terms of the informal communication passage and the associated difficulty in searching and retrieving results from an SMS-based web search engine under its non-standardization. The query is a pre-defined phrase-based translated English version of the SMS. The SMS machine tool normalization algorithm (SCORE) was invented for the query to interface with the best ranked and highly optimized results in the search engine. System results, when compared with a number of open sources SMS translators gave a better and robust performance of translation of the normalized SMS. [4]

III. CORPUS PREPARATION

Corpus preparation plays an important role in overall translation system. NLP techniques use this corpus along with translation techniques to normalize the SMS text into plain text. A parallel corpus contains the SMS abbreviations for translation purpose. A corpus should contain at least 10,000 entries for translation purpose in order to translate the input text properly. Corpus should contain all abbreviations for a particular language for which translation system is to be developed.

IV. APPROACHES FOR TEXT NORMALIZATION

There are mainly three types of approaches are used to translate the SMS text into equivalent plain text which are described in the following section :

A. Dictionary Look up technique

In this technique a parallel corpus is created and results are calculated by comparing the input text with the stored words one by one. This is the easiest and fastest method to obtain the results but works only if the word which is to be translated is present in the database. This approrach fails for those words which can have multiple translations for a single word. For example "2" have three translations which are "two","to" and "too". Dictionary lookup techniques fails to choose best translation for the given input and hence not able to translate. Today most of the websites use this approach for translating the SMS text into plain English text.

B. Rule Based approach

In this approach various rules are to created according to the language model of both source and target language. A lot of experience is required in this domain and user must know all the features of both the languages to make such rules and hence require lot of time and money. Due to changing in styles in writing of the users this approach also fails to translate the text properly.

C. Statistical Machine Translation (SMT) approach

In this approach translation of input text is done with the help of existing translated text. In this approach a large corpus is created which contain input text along with their translated text and output is generated according to the given translated text. This approach works in two phases which are (i) training Phase (ii) Translation phase. In the training phase, various combinations are generated and stored in the system which is used in the second phase. These combinations contain the input text along with the translated text. In the second phase actual translation is done with the help of the combinations generated in the first phase. This approach uses a further approach N-Gram approach to generate the combinations from the input text. This approach is used only upto three grams which provides results with very low accuracy. This approach needs to improved upto six gram to obtain the good results.

V. PERFORMANCE EVALUTION

Most commonly used approaches for text normalization are dictionary look up technique, Rule based approach and Statistical machine translation. Performance evaluation of these techniques is done with the help of Asp.Net.

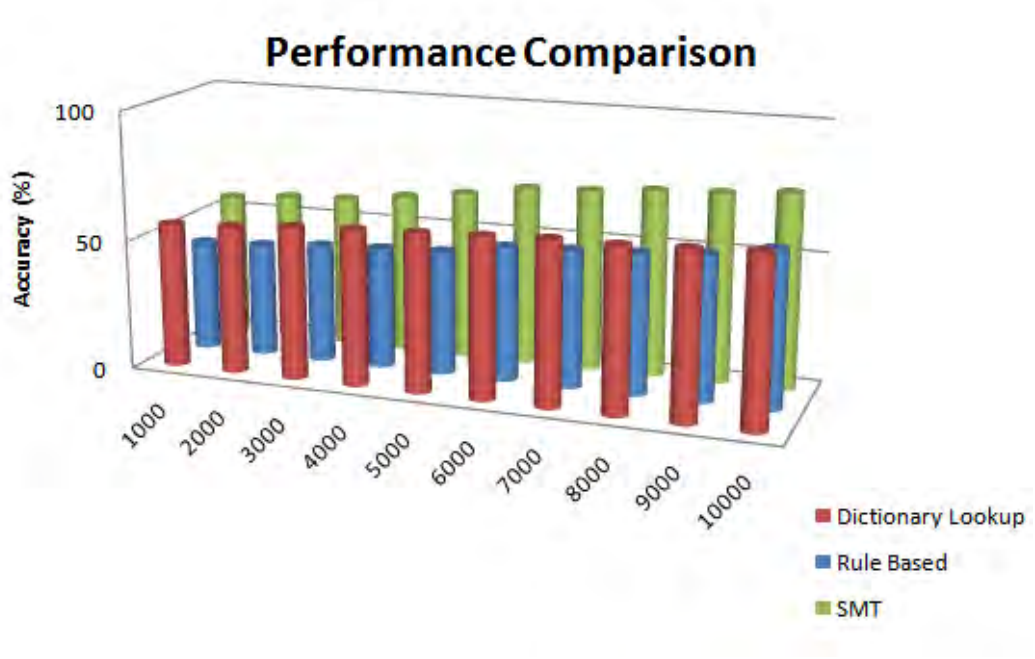
Table 5.1 Comparison between Different Systems:-

Parameter	System[1]	System[2]	System[3]
Accuracy	74.7%	76.23%	74.6%

In table 5.1 we show a comparison of Accuracy of different systems. System[1] means the accuracy of the system as given in [1] reference "Normalization of Text Message For Text-To-Speech" by Deana L. Pennell and Yang Liu. This system is having the accuracy of 74.7%. System[2] means the accuracy of the system as given in [2] reference "A hybrid rule/model-based finite-state framework for normalizing SMS messages" by Richard Beaufort , Sophie Roekhaut , Louise-Amélie Cougnon, Cédrick Fairon. This system is having the accuracy of 76.23%. System[3] mean the accuracy of the system as given in [3] reference "Improving Text Normalization Using Character-blocks based Models and System Combination" by ChenLi Yang Liu. This system is having the accuracy of 74.6%.

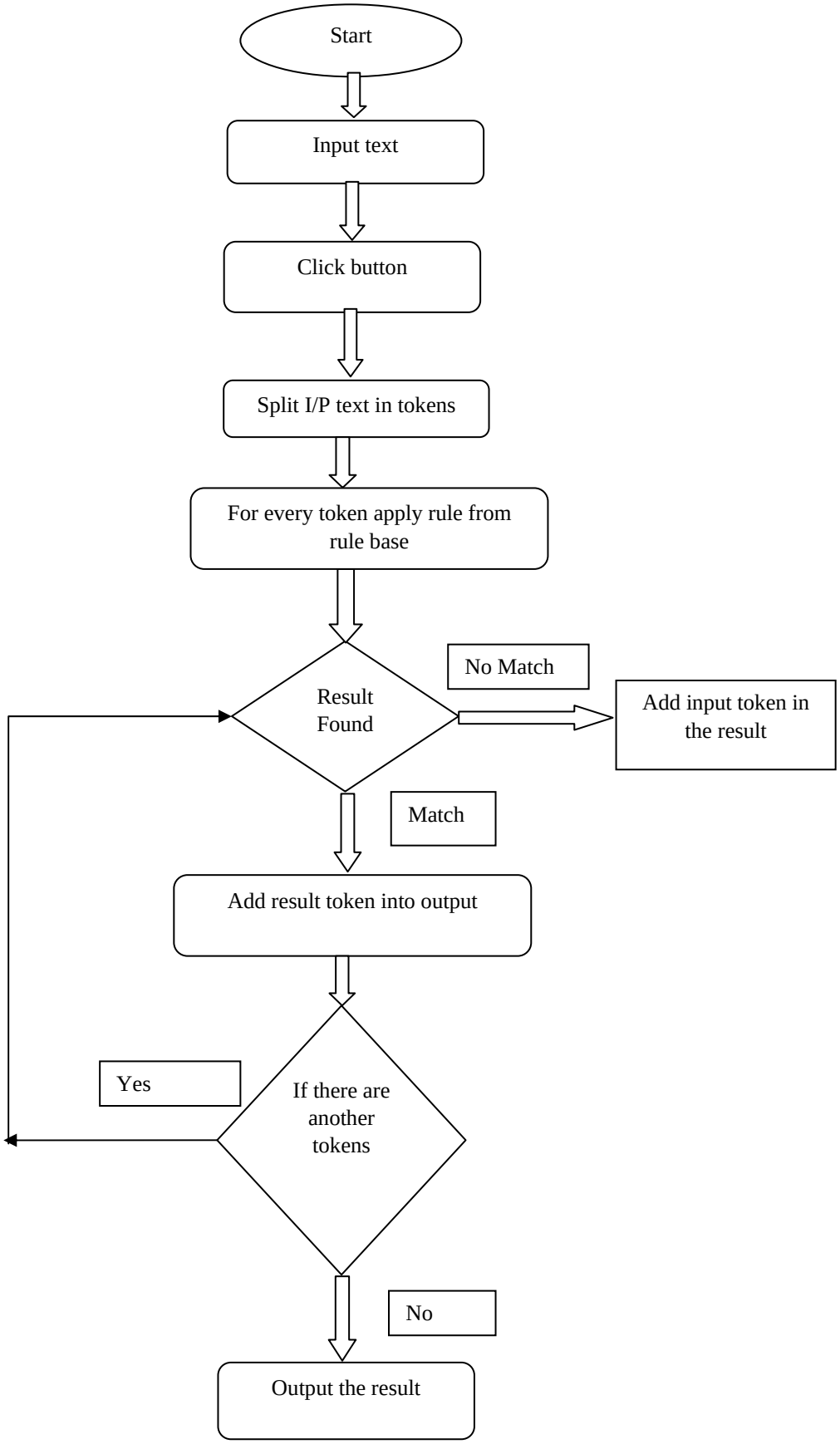
A. Training corpus size

We first started out by measuring the performance of existing systems as a function of corpus size.As can be seen from the learning bar graph in "fig. 1", Accuracy computed for each of the test sets have been increasing with increase in the amount of training data. This shows that our system is still data hungry and we can still hope to get more improvements with additional data.

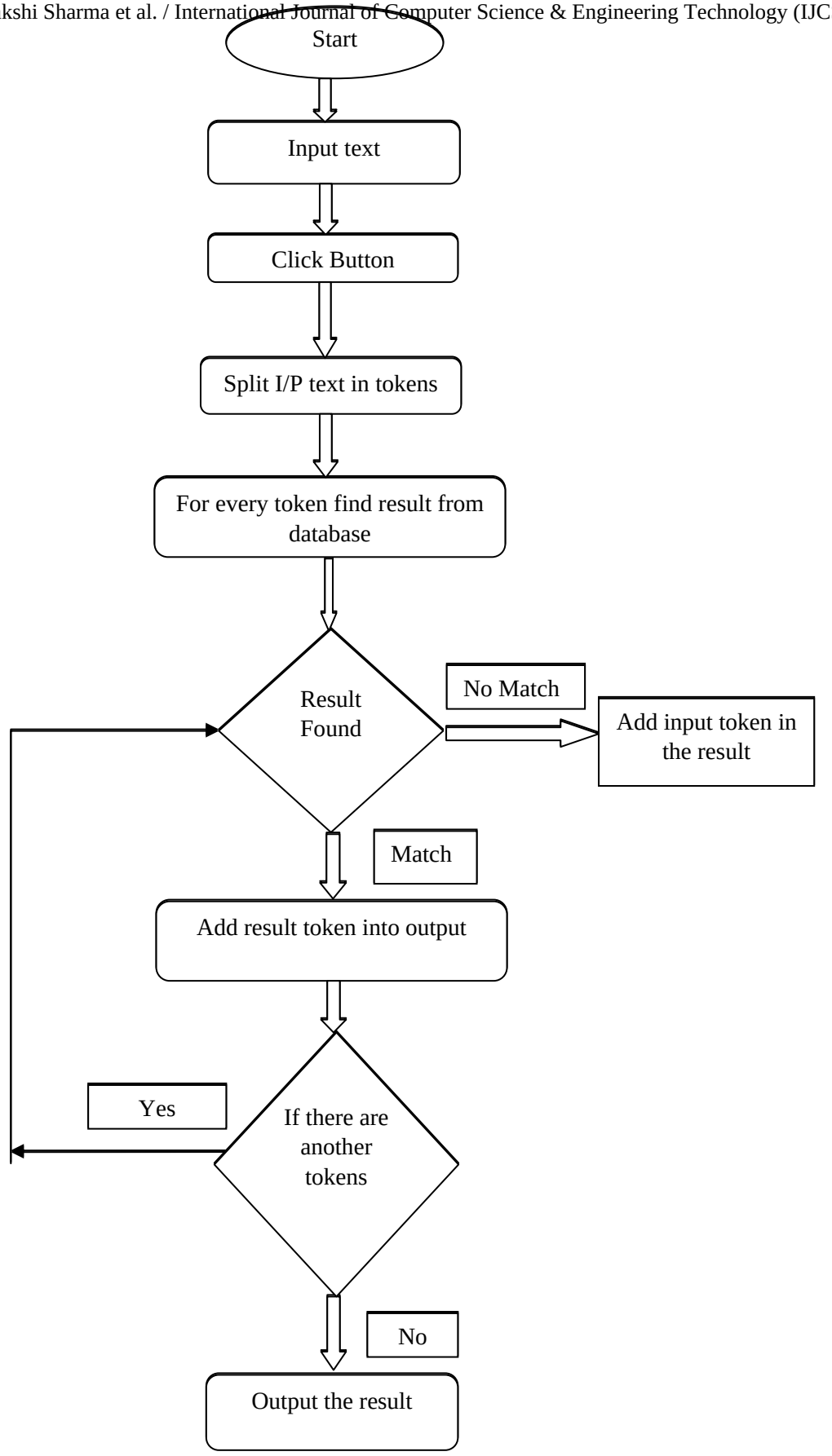


"Fig. 1" :-Performance of various approaches varying the training set size

"Fig. 2" :-Flow chart of Rule based system
"Fig. 3" :-Flow chart of Example based system


Start
Input text
Click button
Split I/P text in tokens
For every token apply rule from rule base
Result Found
No Match
Add input token in the result
Match
Add result token into output
If there are another tokens
Yes
No
Output the result

"Fig. 2"


Start
Input text
Click Button
Split I/P text in tokens
For every token find result from database
Result Found
No Match
Add input token in the result
Match
Add result token into output
If there are another tokens
Yes
No
Output the result

VI. CONCLUSION

In this paper we describe the meaning of the text normalization and various approaches are discussed. Dictionary look up technique is the easiest and fastest method to obtain the results but works only if the word which is to be translated is present in the database. In Rule Based approach various rules are to created according to the language model of both source and target language. A lot of experience is required in this domain and user must know all the features of both the languages to make such rules and hence require lot of time and money. Due to changing in styles in writing of the users this approach also fails to translate the text properly. In Statistical Machine Translation (SMT) approach translation of input text is done with the help of existing translated text. In this approach a large corpus is created which contain input text along with their translated text and output is generated according to the given translated text. We conclude that statistical machine translation is the best technique to obtain good results.

REFERENCES

[1] Deana L. Pennell and Yang Liu, Normalization Of Text Messages For Text-To-Speech , 978-1-4244-4296-6/10/$25.00 ©2010 IEEE

[2] Richard Beaufort , 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 System Combination

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

[5] Zhenzhen Xue, Dawei Yin and Brian D. Davison (2011) , Normalizing Microtext. In proceedings of 25th AAAI.

[6] Raghunathan and S. Krawczyk. 2009. Cs224n: Investigating sms text normalization using statistical machine translation.

[7] J. Chen, et al., "SMS-Based Contextual Web Search," presented at the Mob held '09 Barcelona, Spain, 2009.

[8] Cook, P. and Stevenson, S. (2009). An unsupervised model for text message normalization. In Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, pages 71–78, Boulder, Colorado. Association for Computational Linguistics.

[9] Pennell, D. and Liu, Y. (2011). A character-level machine translation approach for normalization of sms abbreviations. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 974–982, Chiang Mai, Thailand.

[10] Aw, 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.

[11] Choudhury, Monojit, Rahul Saraf, Vijit Jain, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. In Proceedings of the IJCAIWorkshop on "Analytics for Noisy Unstructured Text Data", pages 6370, Hyderabad, India.

[12] Kobus, Catherine and Yvon, Francios and Damnati, Geraldine, "Normalizing SMS: are two metaphors better than one?", Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pages 441–448, Manchester,England.

[13] Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar1, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. International Journal on Document Analysis and Recognition, 10(3):157– 174.

[14] Yu Liping, Pan Yuntao, Wu Yishan, Research on Data Normalization methods in Multi-attribute Evaluation 978-1-4244-4507-3/09/$25.00 ©2009 IEEE.

[15] Llu´ıs Formiga Jos´e A.R. Fonollosa, Correcting Input Noise in SMT as a Char-Based Translation Problem Universitat Polit`ecnica de Catalunya (UPC), Barcelona, 08034 Spain October 31, 2012