Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 1, January 2015, pg.544 – 554

RESEARCH ARTICLE

Text Normalization Using Hybrid Approach

MEENAKSHI SHARMA

Department of Computer Science, Giani Zail Singh PTU Campus, Bathinda, Punjab, India

ermeenakshi89@gmail.com

ABSTRACT

Machine Translation (MT) was an important area of Natural Language Processing that dealt with the translation of one natural language to another language. In this paper we were presenting the research on Translation of short messages to Plain English Text Messages. In today's world where communication over the internet had increased by using various types of websites and another internet applications, short messages were used most frequently for the purpose of communication. These short messages, sometimes were unable to understand by the receiving side and hence need to be translated into the plain English text so that receiver of the message could translate the actual message correctly. Our system used hybrid approach that consist of various approaches like Rule Based Approach, Statistical Machine Translation Approach and Direct Mapping Approach for the purpose of translation from Short message to Plain English Text. Translation of these short messages into plain English text was also known as Text Normalization.

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

INTRODUCTION

Term normalization means to translate the SMS text into the plain English text. It is an important processing step for a wide range of Natural Language Processing (NLP) tasks such as text-to-speech synthesis, speech recognition, information extraction, parsing, and machine translation. The use of normalization in these applications poses multiple challenges. 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 paper 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 ur house den?**" this SMS text can be normalized in the plain English as **"Should we go to your house then ?".**

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]

PROPOSED MEHODOLOGY FOR TEXT NORMALIZATION

Proposed System use hybrid approach to translate the SMS text into its equivalent plain English Text. Hybrid approach consist of mainly three types of approaches are used to translate the SMS text into equivalent plain text which are Dictionary Look up technique, Rule Based approach, and statistical Machine Translation approach. These approaches works in the sequential order i.e. if one approach fails to produce the desired output then another approach will try to produce the result. These approaches 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 approach fails for those words which can have multiple translations for a single word. For example “4” have two

translations which are "for" and "four". 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. Handcrafted rules are made to translate the input text into its equivalent output. While writing a short message users use various symbols and include numerals in their messages which can be only resolved with the help of rule based approach by replacing the symbols or numerals with their actual interpretation.

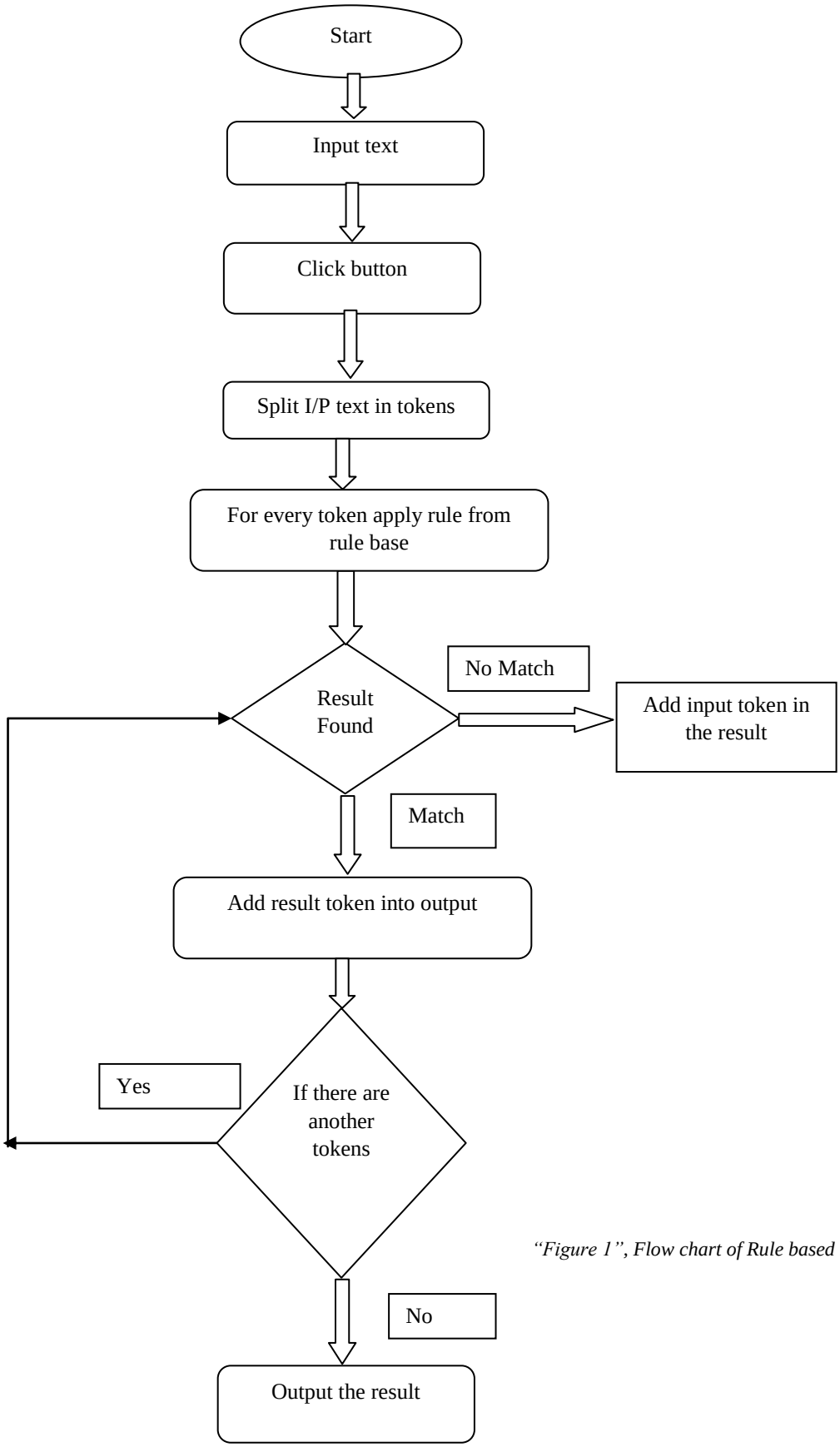
C. STATISTICAL MACHINE TRANSLATION (SMT) APPROACH

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 up to three grams which provides results with very low accuracy. This approach needs to improved upto six gram to obtain the good results.

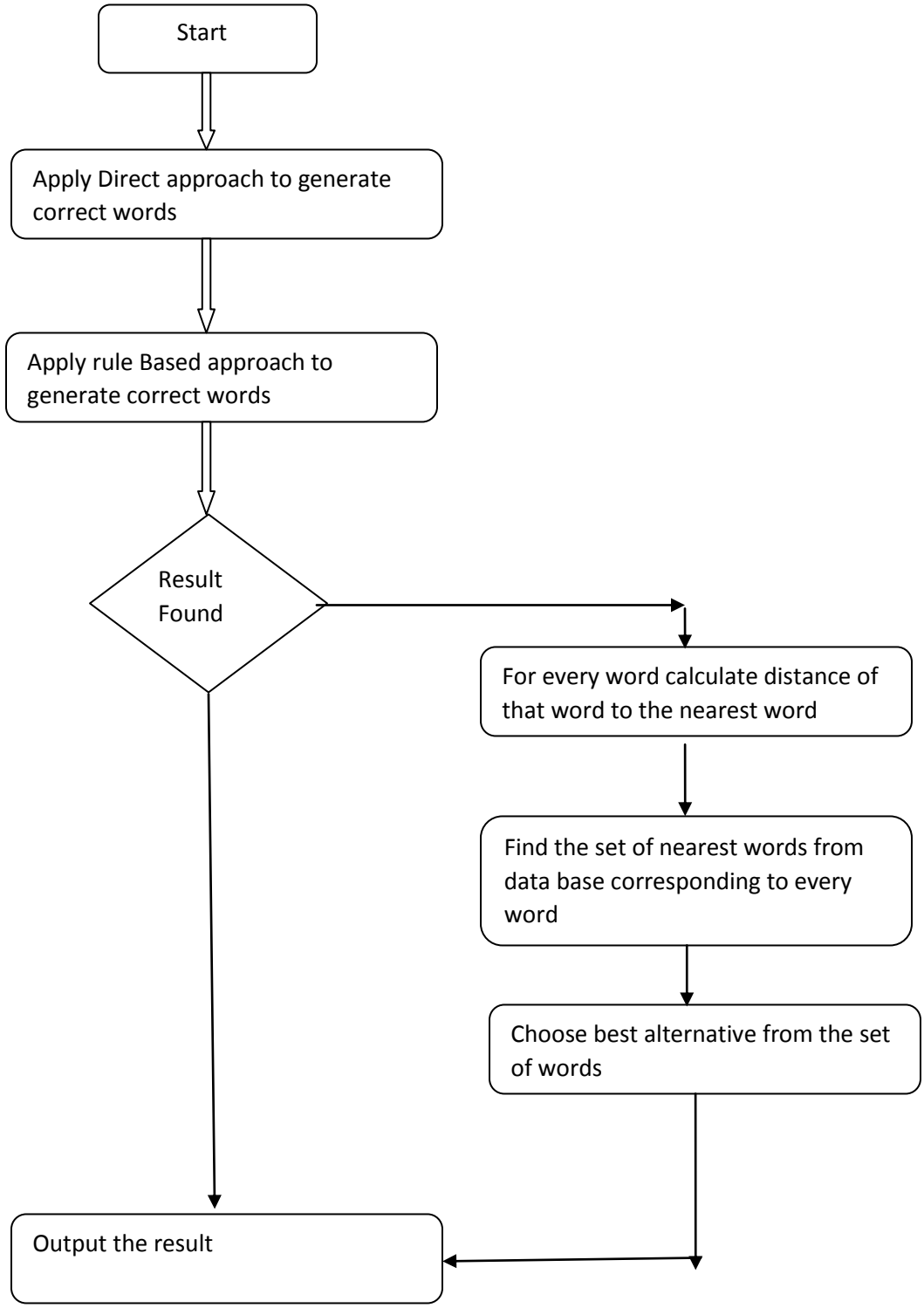
CORPUS PREPARATION

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.

The following are the flowcharts of the proposed system:



"Figure 1", Flow chart of Rule based system



"Figure 3", Flow chart of Proposed system

ALGORITHM

1. Input the short message to be translated.
2. Apply direct approach to translate the input of step1. If result found then go to Step 8.
3. Tokenize the input into six-Grams, Five Grams , four Grams , Three Grams and two Grams. Combine the results from these tokenized words. If result found then go to Step 8.
4. Tokenize the input into individual words and apply rule based approach to extract the result. If results are found for all the words then go to Step 8.
5. Generate the nearest word corresponding to each word whose result is not found in step 4.
6. Add these nearest words into the output of the final result.
7. Align all the output tokens obtained in step 5 and 6 to form a complete sentence according to the input.
8. End

The following tables are generated by our system in Statistical Machine Translation Phase

I. Uni-gram table used for proposed system

Abbreviation	**English**
Dere	There
Wat	What
R	Are
Gng	Going
Wry	Worry
Tnsn	Tension

II. Bi-Gram table generated by our system:

Short Message	**Plain Text**
R u	Are you
Gng 2	Going to
I m	I am
Gud 4	Good for

III. Tri-Gram table generated by our system:

Short Message	**Plain Text**
I m gng	I am going
Wat is d	What is the
Y r u	Why are you
Gng 2 uni	Going to university

IV. Four-Gram table generated by our system:

Short Message	**Plain Text**
Wat r u dng	What are you doing
I m gng to	I am going to
Y r u dere	Why are you there
M gng to uni	Am going to university

V. Five-Gram table generated by our system:

Short Message	**Plain Text**
Sry I m nt cmng	Sorry I m not coming
I m gng to uni	I am going to university
Hlo I m k nw	Hello I m ok now
Plz cme 2 meet me	Please come to meet me

VI. Six-Gram table generated by our system:

Short Message	**Plain Text**
y r u wry I m	Why are you worry I am
I m gng to uni nw	I am going to university now
y shd u wry abt this	Why should you worry about this
I m here 2 hlp u	I am here to help you

VII. The following table shows the results generated by the system :

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

TRAINING CORPUS SIZE

Size of the corpus plays an important role in overall process of translation. System performance automatically increases as corpus size increases. 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.

RESULT EVALUTION

Proposed system contain 12,000 dictionary words for English, 1500 abbreviations and 1000 sample SMS to training the data. Proposed system is tested on 200 entries and accuracy is evaluated to 89.5%.

Result is evaluated for proposed system on the basis of three parameters:

Precision : Precision is the total number of sentences which are converted by our system regardless of semantic of the output. This can be calculated with the help of following formula :

Precision (%) = (No. of Translated Sentences/Total Sentences Tested)*100

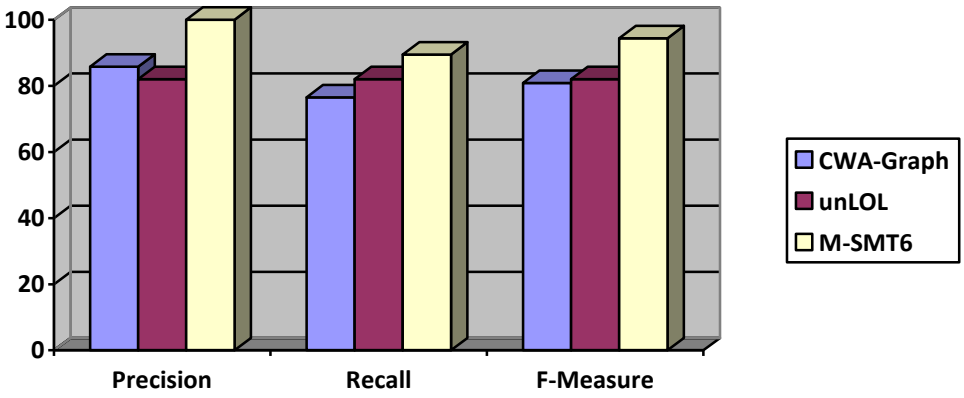
Recall : This is the total number of correct translations out of total translations performed by our system. It can also be termed as accuracy.

Recall (%) = (No. of Accurate Translations/Total No. of Translations) *100

F-Measure=2RP/R+P

VIII. Comparison between Different Systems**:-**

Technique Name	Precision	Recall	F- Measure
CWA-Graph	85.87%	76.52%	80.92%
unLoL	82.09%	82.09%	82.09%
M-SMT6	100%	89.5%	94.4%



"*Figure 4" :-Performance of various techniques varying the parameter values.*

CONCLUSION AND FUTURE SCOPE

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. The proposed system is evaluated to 89.5% accurate when tested on various input sentences. System performance is tested on the basis of three parameters which are overall Accuracy, Precision and F-Measures. In Future work, Rule based approach can be extended further to improve the accuracy of the overall system. Parallel corpus from short message to plain English text can also be improved further to achieve more accurate results. Some other Indian languages like Punjabi or Hindi can also be integrated in the system so that system can become language independent.

REFERENCES

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

[2] Richard Beaufort, et al, 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., et al, A Query-Based SMS Translation in Information Access System.

[5] Zhenzhen Xue, et al , (2011) , Normalizing Microtext. In proceedings of 25th AAAI.

[6] Raghunathan et al. ,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. et al ,(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, et al. (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, et al, 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, et al, 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, et al , "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, et al , 2007. Investigation and modeling of the structure of texting language. International Journal on Document Analysis and Recognition, 10(3):157– 174.

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

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

