A review on SMS Text Normalization using Statistical Machine translation Approach

Sakshi Goyal[1] , Er.charandeep Singh Bedi[2]
[1]M.Tech Final Year Student
Deptt. of Computer Science & Engg. BFCET, Bathinda.

[2]Assistant Professor
Deptt. of Computer Science & Engg. BFCET, Bathinda.

Abstract

Messages often greatly deviate from traditional spelling conventions. 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 used to translate un-normalized text into its equivalent plain text.

Keywords: Text normalization, Direct Mapping, Statistical Machine Translation Approach.

1. INTRODUCTION

Text normalization is a process of translating "text speech" into plain English. Need of normalization is due to these deviations. The small number of characters allowed per text message, the constraints of the small phone, keypads and mostly communication between friends and relatives in an informal register by the service .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. 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 example: "**shd we go 2 yr hme** " SMS text can be normalized in the plain English as

"Should we go to your home" "Y r u gng 2 chd?"
SMS text can be normalized in the plain English as
"Why are you going to Chandigarh?"

2. LITERATURE SURVEY

Karthik Raghunathan, Stefan Krawczyk*:* Investigating SMS Text Normalization using Statistical Machine Translation. In this paper author explore 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. This is followed by a statistical machine translation (MT) approach using off the shelf MT tools. Paper evaluates the performance of the system on three test sets from different sources. It finishes with a discussion about the shortcomings of the system. Firstly, we can try to use a better language model trained on a lot more English data e.g. the Giga-word corpus distributed by the LDC. Secondly, we are right now using an SMS-aligned parallel corpus instead of a sentence –aligned one. It would be worth seeing how splitting multiline SMS into constituent sentences and having a parallel corpus aligned at the sentence level changes the performance of our system. [1] 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 can generate multiple abbreviation hypotheses for each word based on the Classifier's prediction. System will reverse the mappings to enable prediction of English words from the abbreviations. Results show that this approach is feasible and warrants further exploration. Author evaluates 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%. [2] 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%. [3] Chen Li 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. 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%. [4]

3. EXISTING METHDOLOGY

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

3.1. Direct Mapping

In direct mapping approach input text is broken into various tokens and each token is then compared along with database entry and result is evaluated by combining all results obtained from these tokens. This approach will only work if token is present in the database. A large database is required to translate the given un-normalized text into plain English text for this approach.

3.2. 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.SMT is divided into training phase and translation based. During training phase combinations of words are formed from existing data. 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. The key problems in statistical MT are: estimating the probabilities of translation, and efficiently finding the sentence with the highest probability. This system uses n - gram approach which is a combination of the input words .The system has tried upto3-gram approach in the existing system. We can improve the system to 8-grams. Phase

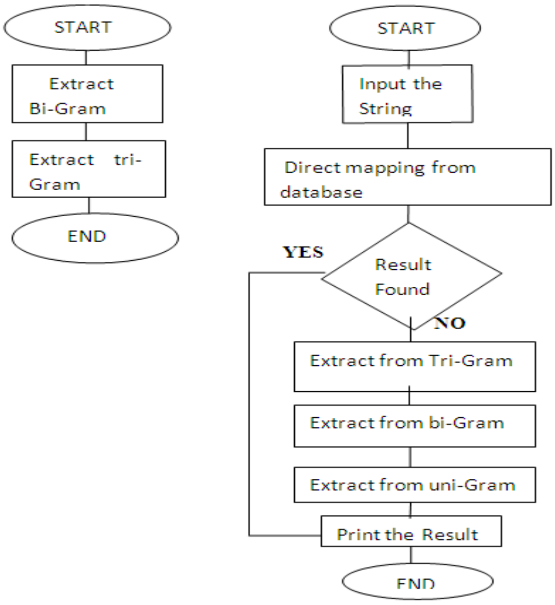

Fig .1[a] Training Phase [b] Translation Phase

4. RESULTS

Table 1: comparison between systems

System	[2]	[3]	[4]
Accuracy	74.7%	76.23%	74.6%

The above system generates the following accuracy which needs improvements. Maximum accuracy of the existing systems is comes out to be approximately 77% which require further improvements.

5. CONCLUSION

In this paper we have present the review on various techniques of translation of un-normalized text into plain English text. We have studied the approaches that lacks in many forms. Hybrid approach which is a combination of all these techniques can be used to obtain more accurate results.

REFERENCES

[1] Karthik Raghunathan, Stefan Krawczyk: Investigating SMS Text Normalization using Statistical Machine Translation.

[2] Deana L. Pennell and Yang Liu, NORMALIZATION OF TEXT MESSAGES FOR TEXT-TO-SPEECH , 978-1-4244-4296-6/10/$25.00 ©2010 IEEE

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

[4] Chen Li Yang Liu, Improving Text Normalization Using Character-blocks based Models and System Combination

[5] Aw, Ai Ti 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.

[6] 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.

[7] 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.

[8] Stolcke, Andreas. 2002. Srilm an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), volume 2, pages 901904, Denver, CO.

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

[10] 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.