International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 08 |Aug -2016 www.irjet.net p-ISSN: 2395-0072

Text Normalization using Statistical Machine Approach

Harpreet Kaur, Er. Jasdeep Singh Mann

Harpreet Kaur ,Student
Er. Jasdeep Singh Mann, Assistant Professor
BMSCE , Shri Muktsar Sahib, Punjab

Abstract: *Natural Language Processing is a research area of an Artificial Intelligence. Various approaches are used to handle text normalization task; in which proposed system uses Statistical Machine Translation approach. Text Normalization is a process in which a SMS text translates into plain English text. Proposed system uses SMT approach which constructs by N-gram and other approaches: Edit Distance, Dictionary Look Up and Direct Mapping. We showed that proposed system achieves 98.4% Precision, 93.3% Recall, and 95.7% F-Measure on the non standard corpus text. Proposed system achieves 90.1% accuracy on both word level and sentence level normalization.*

Keywords: *SMT, Machine Translation, NLP, N-Gram, Edit Distance.*

1. INTRODUCTION: Text normalization is a process of interpreting or transforming informal writing into its standard form. It is an essential processing step for NLP task such as text-to-speech synthesis, speech recognition, information extraction, parsing and machine translation. The main task of normalization is mapping all out of vocabulary (OOV) non-standard word tokens into in-vocabulary (IV) standard forms. It converts raw text into fully grammatical sentence. Now-a-days social text is growing sharply. It includes mobile phone text messages (SMS), comments of social website media for instance facebook, twitter and real time communication data like Gtalk, MSN. Traditional NLP systems poorly process this kind of text, because of its informal nature and misspelled word. For example a text message can be written as follows:
"**wt did u sy**"
Can be translated in the plain English as
"**What did you say?**"

For normalization task we use statistical machine translation approach to normalize the text. Text normalization plays a vital role in any text-to-speech (TTS) system. Text normalization presents numerous difficulties that are not encountered in other domains. This domain is constantly involves; new abbreviations appear frequently and inconsistently. Abbreviations may also have numbers and symbols, which is very unusual in formal text. A typical spell checking algorithm to normalize text messages is mostly ineffective. This is probably because most spelling systems emphasize on single typographic errors using edit-distance, such as a combination of this approach and pronunciation modeling but do not account for the characteristics of text messages. In rapid growth of speech processing applications at low cost, we suggest a system of text normalization systems which are constructed with the support of internet user.

2. CHALLENGES OF NLP: The study of informal text contains features that make it very different from either formal written text. There are few challenges faced by NLP that are discussed below:

Texting abbreviations: The greatest challenge when working with informal text is huge number of out-of-vocabulary words arising from the policy use of abbreviations. These abbreviations have their roots in the first Instant Messaging (IM) systems. As text messaging through the short message service (SMS) on cell phones became increasingly common, people began using abbreviations more frequently to keep their message short i.e. below 160 character limit to save money. These abbreviations pose serious problems for many useful NLP applications. SMS abbreviations have a detrimental effect on machine translation (MT) systems because it would be impossible to store pronunciations in the lexicon for all possible abbreviations. For example –
"MYOB"
can be translated to
"Mind Your Own Business"

Emoticons and Fillers: An emoticon is a facial expression in textual world. In a spoken conversation, humans can use the speaker's tone of voice and visual clues like facial expression, eye movements or lips movements to indicate when someone is being happy, sad, tired, confused or crack a joke etc. In written text, due to absence of these clues its leads to misunderstandings. Emoticons help to provide a voice to the text which leads to better understanding of writer's intention. Emoticons present a difficulty in MT due to a lack of understanding of

how they can be pronounced. Much like texting abbreviations, updated emoticons are constantly being created and multiple emoticons present same purpose.
For example- **:O, :P, ☺** leads to **Surprised, Cheeky, Smile** respectively.

Inconsistent capitalization: In formal English, proper nouns and sentence initial words are always capital. This is not necessarily true in text messages. Capitalization is often used inconsistently even within a single message. Now-a-days, it is quite common to use capital letters to emphasize on particular word or just to make InTeReStInG not at all. Cases of abnormal capitalization indicating emotion or emphasis may actually improve a MT system. A sentence which contains entirely capital letters is generally means that the writer is yelling. On the other hand, inconsistent capitalization can highly grow the difficulty of named-entity-recognition (NER), a task which is relatively easy on formal written English where proper nouns are always capitalized. Lack of capitalization can also effect on sentence boundary detection (SBD) which lead to strange phrasing from the MT system, especially when combined with a lack of punctuation which is very often. Most of the systems don't handle OOV items and related ambiguous inputs.

SMS normalization versus General Text Normalization: General Text normalization leads to Non-Standard words (NSWs) and has been studied in TTS (Sproat et al., 2001) while SMS normalization deals with Non-Standard words (NSWs) or lingoes and has seldom been studied earlier. NSWs for instance digit, sequences, acronyms, mixed case words, abbreviations and so on are grammatically accurate in linguistics. However lingoes, such as **"b4"** (before) and **"bf"** (boyfriend), which are usually self created by young SMS users, are not yet formalized in linguistics. Hence the special phenomena in SMS texts impose a big challenge to SMS normalization. For instance- **"Enuf"** can be normalized as **"Enough".**

SMS normalization versus Spelling Correction Problem: Intuitively, many would regard SMS normalization as a spelling correction problem where the lingoes are erroneous words or non- standard words to be replaced by English words. Researches on spelling correction centralize on typographic and cognitive or orthographic errors and use approaches that mostly model the edit operations using measure distances (Levenshtein 1966), specific word set confusions and pronunciations modeling. These systems are mostly based on character or string without considering the context of message. In addition, author might not be aware of the errors in the word introduced during the edit distance operations, as most errors are due to mistype of characters near to each other on keyboard or homophones such as **"school"** or **"schul"**, **"dreams"** or **"dreamz"**.

3. TEXT NORMALIZATION APPROACHES:

Text normalization consists a numerous of techniques in it to solve any kind of irregularities in SMS, which are discussed as follows:

N-Gram: An N-Gram is a set of consecutive words taken from a string with a length of whatever n is set to integer. Proposed system uses technique up to 8 grams. In SMT these combinations are firstly generate for training purpose to the system and after training is done they are used to normalize the system. Table shows the inputs which train the system for user on various inputs for instance unigram, bigram, trigram, four-gram, five-gram, six-gram, seven-gram, and eight-gram.

Table1. N-Gram combinations

Uni gram	hlo
bi gram	Hlo dr
Tri gram	Hlo dr I
4- gram	Hlo dr I w8
5- gram	Hlo dr I w8 for
6- gram	Hlo dr I w8 for u
7- gram	Hlo dr I w8 for u at
8- gram	Hlo dr I w8 for u at hme

Direct Map: A corpus from short message to plain English is developed which is to be used in the translation process of the system. In this system, a word or a sentence is directly mapped with matched input and generate output.

Dictionary Look-Up: It 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. The accuracy of the proposed system is highly depends upon this phase. If the required word is correct but not is dictionary then it will give incorrect output. E.g.:- Michael, bougainvillea, insomnia etc.

Edit Distance: This technique is used to find the nearest correct possible word from the dictionary to obtain the result. Various suggestions are generated in this phase with respect to the token which is being checked in the

ascending order of their distances. Word distance means the minimum number of operations required to equate the wrong word with the word in the dictionary. E.g. – **"rediculous"** to **"ridiculous"**.

4. PROPOSED ALGORITHM: The proposed system works on the given string using hybrid approach to transform a non-standard text into standard text. Hybrid approach consists of SMT and direct mapping because a single approach is not sufficient to translate any SMS text. A string is input to input box of web interface of the system where numbers of consecutive words are calculated of given string by using N-Gram approach after that on the value of n, apply n- 1 rule. Extracting the table and checking the table according to the value of n. If the desired string is found then it store into a variable. Otherwise replace the value of n by n-1 and check the other tables until n=1, if the desired string not found then direct map approach will work on it , else dictionary look up approach will continue its work, but if string is not found then finally edit distance approach will do its work and generate the output.

Following are the steps of proposed algorithm:

Step 1. Input the string.

Step 2. Calculate the no. of words of the given string.

Step 3. Find the value of n.

Step 4. Apply (n-1) on the value of n.

Step 5. Extract the table and check the table according to the value of n.

Step 6. If String="found" then store the result="tempans"

Step 7. Else replace value of n by n-1; goto step 5 for checking the other tables.

If string! ="found" then repeat the step 5 until n=1 else goto step 8.

Step 8. If string="found" then direct mapping and goto step 6, if string! ="found" then goto step 9.

Step 9. If string="found" then dictionary lookup approach and goto step 6, else if string! ="found" then edit distance approach and then goto step 6.

If string=null then concatenate the tempans else goto step 2.

Step 10. Display the results.

5. CONCLUSION:

The main aim of Text Normalization is to translate the short message into plain English text. The proposed system is based on achieving this aim and comparing the existing and proposed systems. On the basis of number of sentences, the existing systems attained 86.9% of accuracy. The proposed system is compared with the existing systems on the basis of few quality parameters, i.e. precision, recall, F-measure, Accuracy.

Table2. shows comparison of three existing systems with proposed system

System Name	No. of Sentences	Exiting System Accuracy	Proposed system Accuracy
Paul Cook & Stevenson	303	59%	81.8%
Chen Li & Yang Liu	558	86.9%	90.1%
Hany Hassan & Arul Menezes	1000	56.4%	84%

The proposed system achieves 90% accuracy, 98.4% precision, 93.3% recall, 95.7% F-measure when tested on various input sentences. Results of the proposed system shows some improvement over the results of the existing system.

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] Andre Freitas, Sean O'Riain , Edward Curry, Joao C.P. da Silva, Danilo S. Carvalho: Representing Texts as

Contextualized Entity-Centric Linked Data Graphs, In 24th International Workshop on Database and Expert Systems Applications,2013

[6] 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.

[7] Stefan Gerdjikov, Stoyan Mihov, Vladislav Nenchev. Extraction of spelling variations from language structure for noisy text correction. 2013 12th International conference on Document Analysis and Recognition.

[8] Tim Schlippe, Chenfei Zhu, Daniel Lemcke, Tanja Schultz. Statistical Machine Translation based Text Normalization with Crowdsourcing.2013.

[9] Osama A. Khan and Asim Karim. A Rule –based Model for Normalization of SMS Text. 2012 IEEE 24th International Conference on Tools with Artificial Intelligence.

[10] Eleanor Clarka* and Kenji Arakia Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. Pacific Association for Computational Linguistics (PACLING 2011)