ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

Normalization of NSW (Non Standard Words) using DSM model in case of OOV (Out-Of-Vocabulary) words

Dr V.Meenakshi[1], G. Petchiammal Alias Deepa[2], Vidhyalakshmi[3]

Assistant Professor, Department of Computer Applications, R.V. Govt. College, Chengalpattu, Tamil Nadu, India[1]

M.Phil Student, Department of Computer Science, Mother Teresa Women's University Chennai, India[2]

Assistant Professor, Department of Civil Engineering, SRM University, Ramapuram, Tamil Nadu, India[3]

ABSTRACT: In a social web services like Twitter enables any user to send his or her own messages as information and also can able to read the information as tweets. A short message type Tweets are usually of some 130 or 140 characters around or it may be less than that and is shared by many. Tweets may not be in its purest form and it is a raw form but provides useful and important information and tweets it is also vast. It is primary and foremost thing to identify the noise around the data and the data that is redundantly occur. In Tweets there may be OOV (Out-Of-Vocabulary Words) which may not by recognized easily. This type of OOV words are coming under NSW (Non Standard Words). NER (Named Entity Recognition) words are vast number of words recognized by the social media. In a small and static data sets it may create a major problem for the understanding of the Tweets. But for a data of large data sets if the OOV is sparsely occur then the percentage of risk may not be that much scalable but if the OOV occur redundantly then an efficient method is to be devised to get the correct Tweets. Here in this paper a novel method named as DSM (Domain Specific Model) has been used to identify the token whether the tokens in the text (Tweets) are correct-OOV or incorrect-OOV. This model has been given to correctly identify the OOV words and measures are provided to reduce the percentage of risk.

KEYWORDS:Tweets, OOV (Out-Of-Vocabulary) words; NER (Named Entity Recognition); NSW (Non Standard Words), DSM (Domain Specific Model).

I. INTRODUCTION

Short and Quick text messages and comments are the most popular as far as Twitter and Face Book social Websites are concerned. These provide strong and popular communication forms in recent times. Texts in these includes abbreviations, short words, and mis-spelled words. Hence these words are coming under NSW (Non Standard Words). The main example NSW are coming under length limitation of words. Rather providing full length words their short forms are used. The user's main aim is to post message in a precise manner and they need the words to be in a stylish way. This naturally pose a serious problem for the NLP (Natural Language Processing) Domain. The exact word may not be in the NLP domain. Hence sound techniques are needed to improve and enhance the language processing capabilities and to improve the data in the social media. For this we need a strong normalization procedure for the inclusion and the exclusion of OOV words into NSW words and to provide a corresponding IV (In-Vocabulary) word or Standard words.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

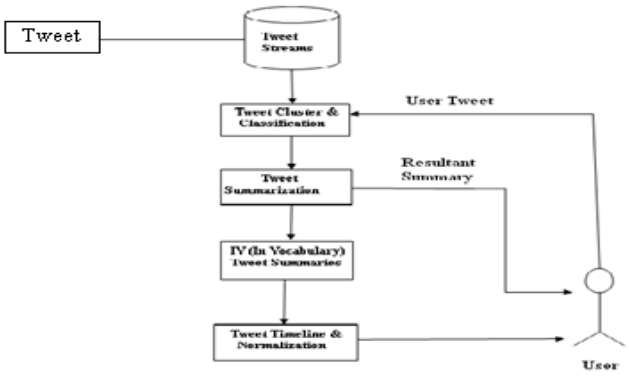

Fig1.1 Architecture Diagram of Tweet summary, Classification and Normalization

Fig 1.1 depicts the architecture diagram of the tweet of any social web-data. The tweets are classified as Tweet streams and they further classified as Clusters and Classifications. Then, Summarization process is done based on IV and OOV. Finally, the resultant summary is sent to the user. The normalization processing tasks now-a-days have been received an increasing attention in the social media language like Twitter and Face Book.

However, many of the previous works on normalization concentrated on tokens which they know are NSW that need normalization. Different modes of handling these tokens are taken by many researchers. Any token in an informal text can be categorized into three different classes namely IV words which are recognized as which naturally belong to the domain and the second class is Correct-OOV which we can say is a Out-of-Vocabulary word but can be taken as NER(Named Entity Recognition) which may includes words like SUDOKU a game, NOODLES a eatable. Then the lastly classified tokens are incorrect-OOV which may be an unknown token. Our experiment is to incorporate all these types of tokens and come out with a best possible result.

GNU spell dictionary (v0.60.6.1) is used for our experimental work to determine a token whether it is an OOV or not. URLs and Hash-tags are not been considered as OOV. Predefined dictionary defined for the internet is used to filter the incorrect-OOV words. The work has been broadly classified under two-way process. From the first step we label a token as IV token or OOV token. Hence every token is classified into two categories namely whether the token is in the domain or not. In the second process those tokens which come under the category OOV are processed to filter as NER or not. Thereby we identify NSW of the given text. Any text which with lexical features are focused to identify whether it comes under the class of regular English word or not. Mostly the correct-OOV includes location, person names, games, things etc., these are still included as standard words. Hence it is our aim to find out the non-standard words. Here for any OOV, we specifically used a Domain Specific Model (DSM) to distinguish between correct-OOV and incorrect-OOV. In the DSM, after checking the two way classification of IV and OOV, DSM is mainly used to identify the correct-OOV and incorrect-OOV.

Lexical features of each and every token is used to identify correct English words. By this way, the IV tokens of the given text are identified. OOV words does not comply with any words of the formation rules of correct English word. In this paper, (1) weidentified IV and OOV by a NSW detection model by normalization information of the OOV words. (2) A data set with new NSW is created for every particular domain in addition to the existing NER. (3) DSM is proposed to clearly identify the token to which it actually belongs to and thereby detecting NSW. This model is run at the same time using social media data and finally (4) we demonstrated the effectiveness of our proposed DSM that provides a means of improving the effectiveness of the system.

II. RELATED WORK

Aw, et al[2005][1] offered an input normalization for an English to Chinese SMS Translation System. A phrase based statistical model has been is provided for sms text normalization by Aiti Aw, et al [2006][2]. He offered an approach to solve the irregularities thereby normalize SMS texts using MT(Machine Translation). They used a phrase based

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

statistical MT model. Their method used a corpus which includes 5000 sentences and they are able to score 0.80702 in BLEU than the baseline score of 0.6958. Deana Pennell and Yang Liu, 2010[3] worked on Normalization of text messages for the text to speech. Deana Pennell and Yang Liu, 2011[4] also offered a character level machine translation approach for the normalization of SMS abbreviations.

Han and Baldwin (2011) in his work did a pilot research on Non Standard Word detection. He used a straight forward method in which he used a dictionary in order to classify a token into Out-of-Vocabulary (OOV) words , In-Vocabulary (IV) words and, and he treated all the OOV words as Non Standard Words. For example, tokens like 'Maggie', 'SUDOKU' (a game name) and 'SUDOKU' will be considered as NSW. But these words cannot be considered as OOV and does not need any normalization. Hence Han and Baldwin[2011] [5]named these OOV words as correct-OOV, and they like to name the OOV words which need normalization as incorrect OOV and ultimately their work is based on formulating a normalization for ill OOV. We like to follow the naming convention of using the two terms of our study as correct-OOV, and incorrect OOV.

Language processing is the major thing in processing the social media web contents. To improve the language processing there are many ways of improving the performance of social media data. One such approach is to normalization leverage technique which automatically converts the NSW words (Sonmez and Ozgur, 2014 [6], Li and and Liu, 2014[7], Liu et al.,2012[8] , Pennell and Liu, 2011[9], Cook and Stevensoon2009[10]) into its corresponding standard words.Many increasing research works havebeen done which integrated lexical normalization withNLP tasks. Kaji and Kitsuregawa(2014)[11]did researches in which they combined lexical normalization, POS tagging and word segmentationon Japanesemicroblog. With the help of character level and word level features Kudo et al.,2004[12], Neubig et al.,[2011][13] did word segmentation and Part Of Speech (POS) tagging in Japanese. Li and Liu 2015[14] did the same POS tagging with text normalization for English. Wang and Kan(2013)[15] also came with the proposal of joint ill-OOV wordrecognition and the word segmentation in Chinese Microblog.But in this method, ill-OOV wordsare only recognized but not normalized. Hence,they did not investigatedthe method of exploiting theinformation derived from normalizationto increase the word segmentation accuracy. Liuet al. (2012) [8] also studied the problem of (NEN) Named Entity Normalization for tweets. They hadan idea of graphical model to conduct both NER and NEN simultaneously on multiple tweets. This work involved text normalization but focusedon NER task. Kucukand Steinberger (2014) [16] worked on Turkish Tweets and adapted NER rulesand other resources to better fit the Twitter language. They relaxed its capitalization constraint, expanded itslexical resources based on diacritics, and finally used anormalization scheme on Tweets. In their work they got positive effect on overall NER performance.

Rangarajan Sridhar et al., (2014)[17] managed the SMS translation task into normalization coupled with translation. Hence they exploited the bi-text resources, anddid the normalization approach using words learned throughneural networks.In this study, we proposed a new methods which include DSM model to effectivelyinvestigate theinformation of OOV words and the normalization of NER task. This study systematically evaluates the effect ofOOV words and normalization on the NER in socialmedia data.

III. PROPOSED METHOD

The work has been broadly under two-way process. Initially, we label every token as IV token or OOV token based on their feature identification in the domain Viterbi decoding is used for the feature identification. Hence every token is classified into two categories namely whether the token is in the domain or not. So, it is possible to classify the given token as IV or OOV. In the second process those tokens which come under the category OOV are processed to filter as NER (Named Entity Recognition) or not.

NER are specialized words which are identified as Standard words of any Domain which preferably are names of the cities, names of the games, or any product name, location name, person name etc., in this stage, we could clearly identify NSW of the given text. Any text which with lexical features are focused to identify whether it comes under the class of regular English word or not.

Mostly the correct-OOV includes location, person names, games, things etc., Thesefeatures are still included as standard words. Hence it is our aim to find out the non-standard words. Here for any OOV, we specifically used a Domain Specific Model (DSM) to distinguish between correct-OOV and incorrect-OOV. In the DSM, after checking the two way classification of IV and OOV, DSM is mainly used to identify the correct-OOV and incorrect-OOV.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

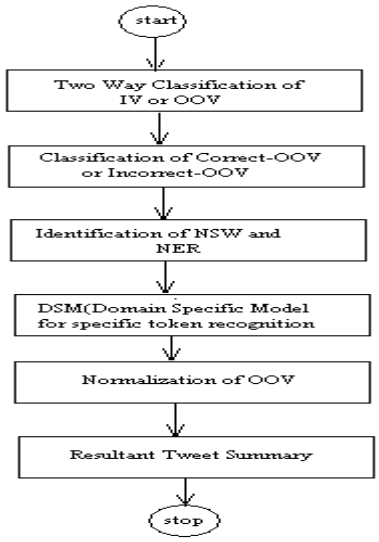

Fig 3.1 Work flow of a Tweet incorporating DSM

Fig. 3.1 shows the diagrammatic flow of the Tweet incorporating the proposed model DSM. Lexical features of each and every token is used to identify correct English words. By this way, the IV tokens of the given text are identified. OOV words does not comply with any words of the formation rules of correct English word. In this paper (1) we identified IV and OOV by a NSW detection model by normalization information of the OOV words. (2) A data set with new NSW is created for every particular domain in addition to the existing NER. (3) DSM is proposed to clearly identify the token to which it actually belongs to and thereby detecting NSW. These model is run at the same time using social media data and finally (4) wedemonstrated the effectiveness of our proposed DSM that provides a means of improving the effectiveness of the system.

Non Standard Word (NSW) Detection: The terms or words which indeed need normalization are the NSW words. After categorizing every token of text as IV and OOV the task of categorizing correct-OOV and incorrect-OOV comes into effect. Hence any token can have only three types of classification namely IV, correct-OOV, and incorrect-OOV.

Table 1. Lexical Features

S.No	Features
1	Word identity
2	Capitalized first character
3	Length of the token
4	Number of Vowels
5	Number of Consonant characters of the token
6	Length of longest consecutive vowel character chunk
7	Length of longest consecutive consonant character chunk
8	Number of consecutive same character
9	Character level probability of the token based on a language model

The normalization needs lexical features to be identified for every token. Some of the lexical features are given in the table 1.The task of normalization is done for the incorrect-OOV tokens. For this experimental study, we consider only single token OOV. Here in this paper, we use GNU spell dictionary (v0.60.6.1) to determine whether a token is OOV. Twitter mentions (e.g., @twitter), hash-tags and URL s are excluded from consideration for OOV. Dictionary lookup of Internet slang2 is performed to filter the incorrect-OOV words whose correct forms are not single words

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

Normalization Features: Normalization features include investigating if any of the individual candidate list has similar token like this and at the same time finding out how many similar candidate list are having the same token. i.e., number of similar tokens of the list. Normalization checks the similarity between the token of the first candidate **w** and this candidate's token **t**.

This is calculated by $\frac{\text{Longest Common String}(w,t)}{\text{Length}(t)}$

Normalization also investigate each of individual lists top 10 list that contains the token of this and also consider the maximum number of lists which have the same token of the candidate.

Dictionary Feature: Dictionary Feature for any token is categorized as IV or OOV for any given dictionary of words (3-way classification).

Features that are in NSW detection includes features like lexical features, normalization feature and dictionary features

NER (Named Entity Recognition) Detection:

This type of entities includes person name, location name, game name, organization name etc., Basic features that includes any NER system has

1) Lexical features which include N-gram Model. i.e., Unigram, Bigram and Trigram.
2) POS features which includes Unigram, Bigram and Trigram
3) Information about the token's capitalization
4) Dictionary Characterization
5) Predicted NSW label etc.

Domain specific Model(DSM): Here we propose the method of forming a specific domain which pre-includes words or token which may probably occur in a particular context. This can also be called as context based domain which pre-requires all sorts of NER that are probably a correct token. In this way a NSW word can be checked for its position in the domain specifically designed for and is treated as correct-OOV. The check here is done for the OOV's presence inside the domain or outside the domain. This type of leverage method provides a better tweet and performance is also increased. Here we have to say POS (Part Of Speech) tags that are designed for media data or media text.

Summary generation: The summary finally is to be generated for each and every topic of the tweet which should be ranked accordingly. Top ranked tweet of the subtopics are also selected. So both at tweet level and sub-topic level the corresponding summary is generated. Scores of similar tweets are ranked and the high similarity tweets are chosen accordingly.

IV. DATA SET AND EXPERIMENT

Data Set and Experiment: The NSW detection model is used for training the data which had more than 3000 Twitter messages in which 2577 unique pairs of NSW and its corresponding standard words. The data used for training have different normalization model. This is been labeled as data for the dictionary which are to be used for NSW detection. Tokens are separated from the Twitter messages. Out of the total tokens, 3122 tokens are given the label as incorrect-OOV, 1627 tokens are given the label as correct-OOV. The remaining 37,549 tokens are IV words.
The entire sets of tokens are separated into two sets to evaluate the NSW detection. First set of tokens include 634 tweets. Every tweet has at least one incorrect-OOV and its corresponding correct word. This we call it as Test set1. The other set contains 756 tweets for the detection of NSW. We call it as Test set 2. We train all 3122 tokens for the POS model to find the part of speech and the corresponding tokens. SRILM (Stolcke 2002) [18] is used for constructing the Language Model (LM) for the tweet messages. This toolkit produces character level LM which generates LM features

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

for the detection of NSW. Cheng Li and Lang Liu, 2015[19] investigated about improved named entity recognition in Tweets via Detecting Non standard words.
In the next phase after the detection of all token using NSW detection, label every OOV as incorrect-OOV. These labeled OOV's are treated for its correctness based on its presence in the DSM domain. DSM domain includes words or tokens of predefined any specific contextual content. Here the textual identification based on different features are used to identify the correct token. By this method, all OOV's are identified as correct-OOV or incorrect-OOV. Then the percentage of correctness are identified and compared with all other previous methods.

Experiment Result:
NSW Detection Results: For the NSW detection model we have compared the proposed system using DSM on two test sets and conducted different experiments and we investigated the effectiveness of various features. Categorization of words using dictionary is treated as baseline for our model.

Table 2: NSW Detection Results

System	Test Set 1			Test Set 2		
	R	P	F	R	P	F
Dictionary	88.53	70.77	72.78	66.78	68.76	68.56
Two-Step	82.22	86.56	83.34	58.55	88.98	72.32
3-way	87.89	83.77	89.91	71.56	90.32	82.25

Table 2 shows the results of the three NSW detection systems. R,P,F respectively is used for Recall, Precision and F value for the incorrect-OOV class which acts as the evaluation metrics. With the results of the table we could see both the two-step and the 3-way classification modes leverage the features described are not given expected reliable result and also we could see other normalization feature information may be more stable.

Table 3: NSW Detection Results

System	R	P	F
3-Way Classification	59.43	71.21	63.21
DSM inclusion	60.32	73.11	66.22

From this table3, we could see when using our DSM system, adding the NSW label features has a very significant improvement compared to other basic features. The F value of 66.22% by using all features is even higher than state-of-the-art performance.

V. CONCLUSION

In this paper, we proposed a novel approach for the detection of NSW. Our proposedNSW detection system leveraged the normalizationinformation of OOV and also other useful lexicalinformation. Hence, this method makes the lexical normalization a complete applicable process. The experimental results shows bothkinds of information could help improve the predictionperformance on two different data sets. Furthermore,we applied the already predicted labels as the additionalinformation for NER task. In thistask, we proposed a novel decoding approachto label each token's NSW and the NER label in tweet at the same time.
The experimental results also demonstrate that NSW label has a very significantimpact on the NER performance and our proposed DSM method improves overall performance on both tasksand outperforms the best previous results in the NER. As our future work, we like to pursue the research work on different numberof directions. First, we like to plan how toconduct NSW detection and the normalization procedure at the same time.We too like to try for a joint method which simultaneously will train NSW detection and NERmodels, rather than the existing combining models in

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 8, August 2016

decoding. And we also like to investigate the impact of the NSW and normalization on the other Natural Language Processing tasks suchas parsing of data in social media data sets.

REFERENCES

[1] Aw,A.T., Zhang ,M., Fan, Z.Z., Yeo, P.K. and Su, J., " Input Normalization for an English-to Chinese SMS Translation System" MT Summit - 2005.
[2] Aiti Aw, Min Zhang, Juan Xiao and Jian Su., "A phrase-based statistical model for sms text normalization" In Processing of COLING/ACL.
[3] Deana Pennell and Yang Liu., "Normalization of text messages for text-to-speech" In ICASSP, 2010.
[4] Deana Pennell and Yang Liu., "A character-level machine translation approach for normalization of sms abbreviations" In Proceedings of IJCNLP.
[5] Bo Han and Timothy Baldwin., " Lexicalnormalisation of short text messages Maknsens a #twitter" In Proceeding of ACL,2011.
[6]CagilSonmez and ArzucanOzgur., " Agraphbased approach for contextual text normalization" In Proceedings of EMNLP, 2014.
[7] Chen Li and Yang Liu., " Improving text normalization via unsupervised model and discriminativereranking" In Proceedings of ACL,2014.
[8] Fei Liu, FuliangWeng, and Xiao Jiang., " A broad-coverage normalization system for social media language" In Proceedings of ACL,2012.
[9] Deana Pennell and Yang Liu., "A character-level machine translation approach for normalization of sms abbreviations" In Proceedings of IJCNLP, 2011.
[10] Paul Cook and Suzanne Stevenson., " An unsupervised model for text message normalization" InProceedings of NAACL,2009.
[11] NobuhiroKaji and Masaru Kitsuregawa., "Accurate word segmentation and pos tagging for Japanese microblogs: Corpus annotation and joint modeling with lexical normalization" In Proceedings of EMNLP, 2014.
[12]TakuKudo, Kaoru Yamamoto, and Yuji Matsumoto., "Applying conditional random fields to Japanese morphological analysis" In Proceedings of EMNLP, 2004.
[13]GrahamNeubig, Yosuke Nakata, and Shinsuke Mori., "Pointwise prediction for robust, adaptable Japanese morphological analysis"In Proceedings of ACL, 2011.
[14] Chen Li and Yang Liu., "Joint POS tagging and text normalization for informal text" In Proceedings of IJCAI, 2015.
[15] AoboWang and Min-Yen Kan., "Mining informal language from Chinese microtext: Joint word recognition and segmentation" In Proceedings of ACL, 2013.
[16] DilekKucuk and Ralf Steinberger., " Experiments to improve named entity recognition on Turkish tweets"In Proceedings of Workshop on Language Analysis for Social Media (LASM) on EACL, 2014.
[17]Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, and Ron Shacham., "A framework for translating SMS messages" In Proceedings of COLING, 2014.
[18] AndreasStolcke., "SRILM-an extensible language modeling toolkit"In Proceedings International Conference on Spoken Language Processing,2002.
[19] Chen Li and Yang Liu., " Improving Named Entity Recognition in Tweets via Detecting Non-Standard Words" In Proceedings of the 7th International Joint Conference on Natural Language Processing, pp. 929–938, 2015.