ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, Vol 8(27), DOI: 10.17485/ijst/2015/v8i27/81666, October 2015 ISSN (Online) : 0974-5645

Toward Normalizing Romanized Gurumukhi Text from Social Media

Jagroop Kaur and Jaswinder Singh*

Department of Computer Engineering, Punjabi University, Patiala - 147002, Punjab, India; dr.jaswinder@pbi.ac.in

Abstract

Roman characters are used to write Indian language text on social media like facebook and twitter. Processing this text for NLP applications is not a trivial task. This text needs to be transliterated as well as conversion to canonical form. This paper discusses the various issues involved in normalizing such text in the domain of Punjabi Language. An algorithm is proposed to normalize Punjabi language text which is written using roman script. The proposed algorithm tries to find out all possible combinations and then filter using n-gram language model.

Keywords: Text Normalization, Transliteration, Social Media Text

1. Introduction

Among the gifts of God for the human being, the most beautiful is the ability to communicate. With the evolution of mankind, the way of communication also evolved. People start communication through different medium like written form, spoken form and using sign etc. Different kind of languages had been developed for effective communication among the masses. In multilingual societies like India, a very common phenomenon is mixing of multiple languages for expressing one's ideas. In earlier days, mixing of languages is restricted in spoken form. The invasion of electronic media and internet simply intensified the phenomenon of mixing multiple languages in written form also. Now a days, social site's platform like facebook, twitter etc. contain lot of text in two or more languages. These texts contain information in abundance for example the text may be about some product or some event. By analyzing such text one can find out the usefulness of some product or event. Recently the research community in Natural Language Processing generated interest in analyzing code mixed text and start developing algorithms to process such type of text.

One of the major issues of text from social media is un-normalized text. This text is written in casual way which often do not confirm to rules of grammar spelling and punctuation. It is something beyond simple spelling correction. The text from social media like facebook, twitter etc. is appearing in various forms. People intentionally type non-canonical data. These text are entered through different devices like mobile phones, tablets etc. Due to lack of comfortable input medium, people tend to write in short form to save character or keystroke. Various issues that one need to address during normalization are slangs, acronyms and short forms,omission of punctuation marks or stylish use of punctuation marks, phonetic spelling, misspelling etc. For example, repeating letters or punctuation for emphasizing and emotional expression such as "goooooodmorniiing". Using phonetic spelling in a generalized way or to reflect a local accent; such as 'wuz up bro' (what is up brother). Eliminating vowels such as 'cm to c my luv'. Substituting numbers for letters such as '4get' (forget), '2morrow' (tomorrow), and 'b4' (before). Substituting phonetically similar letters such as "phone" (fon). Slang abbreviations which usually abbreviates multi-word expression such as LMS (Like My Status), idk (i do not know), rofl (rolling on floor laughing). For a language other than English, the problem escalated due to use of roman alphabets to represent words in language under consideration. For example a text in Gurmukhi script but written using roman alphabets found on a social site facebook as follow:

* Author for correspondence

"J eh gal sahihaitaabhut hi vdiyahai"

"1950 vich ta remote v discover ni Hoya c…Eh v 1980 vichAyaaa.Gapplagdiaamainu ta"

Normalising such text involves much more problems than discussed above. Next section discusses the issues involved in normalizing such text followed by a section on literature survey. Section 4 describes an approach to transliterate and normalize Gurumukhi text written using Roman alphabet. Section 5 discusses the evaluation and results.

2. Issues

On social media, people generally use roman alphabets for writing Punjabi, Hindi as well as English. This is due to the fact that they are not aware of Unicode based fonts and they even don't know how to type in languages like Punjabi or Hindi through their keyboard or through their mobile/tablets. This poses a first difficulty in the analysis of data from social media i.e. identification of language of text. e.g. in a comment posted on facebook page.

"**Meanwhile MST** de vichhameshakudiyan nu jyada **Marks**kyonmilde a? C**lick here to see WHY**" In this sentence bold words are from English Language and normal words are from Punjabi Language. In some cases, the sentence contains words from three languages viz. Punjabi, Hindi and English. For the current task, it is assumed that the domain of input text is Punjabi language but written using roman alphabets. Following section discuss the issues of normalizing such text. Normalizing here means to convert text written in roman alphabet to text in Gurmukhi script. The task can be treated as a combination of transliteration and normalization task. The system need to find out the appropriate canonical form from the text in written in roman alphabets. Some of the issues are:

2.1 Different Way of Writing

While using roman script, every user write with his own way. For example the word "ਜਰੂਰੀ" can be written in various ways like zaroori, zaruri, jaroori, jaruri, zaroory etc. So system need to identify that they all belongs to same canonical form in target script.

2.1.1 Text Shortening

The text from social media makes frequent shortening of long words. For example the word "ਕਰਦੇ" may be written as "krde" in roman form. But user shortened it by omitting some characters like "krde". Some other examples are:krn (ਕਰਨ), krta (ਕਰਤਾ), ght (ਘੱਟ) etc. Generally, phonetic equivalent from roman text that represent same kind of corresponding sound, are selected to make words.

2.1.2 Phonetic Similarity of Spellings

Due to phonetic typing some words share the same surface form. For example the word "banda" can bemapped to "ਬੰਦਾ" or "ਬਣਦਾ". This problem further escalated if three languages are involved. People belongs to Punjab generally use mix of Punjabi, Hindi and English. Thus a word may map to hindi word or Punjabi word. As an example, "to" is a word in the three languages: in Hindi it is 'ਨੋ' and in Punjabi it is 'ਤੋ' and in English it is 'to'. Ambiguity resolving module and context feature is required for such type of problems.

2.1.3 Dialectical Variations

As social media text is informal way of communication, people generally uses the dialectical variation of words. e.g. the word "ਆਖਦੇ" has dialectical form "ਆਹਂਦੇ". In roman script, people use the word "ahnde" which need to be transliterated as well as further converted to the canonical form.

2.1.4 Typing Errors

A very common problem is error due to typing. As the words are entered through text area which supports English language, the auto correction feature of input field sometime changes the letters automatically. Consider following sentence:

"Par Mere sareernal koi faltucehjnhai jure hoie"

Here the word "cehj" which represent "ਚੀਜ਼" is wrong. The correct version is "chej". Similarly, the word "nhai" should be "nahi"which represent "ਨਹੀਂ".

2.1.5 Intentional Deviations From Standard Orthography

The text contains words like kmaaaaalll!!! (ਕਮਾਲ) hanjiiiii!!! (ਹਾਂਜੀ) oh teriiiiiii!! (ਓਹਤੇਰੀ). The repeated letters are for showing emphasis. Identifying these words and correcting them also intensify the problem.

2.1.6 Non-Linguistic Sounds

Writers express their happiness or sadness using

lexical terms that are not the part of language. E.g. to express happiness people write "hahahaha", "HeHeHe", "Buraaahhhhhhhh". Such type of text must be handled adequately.

2.1.7 Non Standard Abbreviation

Writers use non-standard abbreviations like 22 g for Baiji (ਬਾਈਜੀ), 7nam for Satnam (ਸਤਨਾਮ). This also varies from person to person. Identifying such abbreviations added extra difficulty to the problem.

2.1.8 Multiword Tokens

Some time users enterMultiple words as single token. For example "gm" for "gol market".

2.1.9 Creative Use of Punctuation

People use punctuation creatively on social media. e.g. :-) used for represent happiness and :-((for sad mood.

2.1.10 Little contextual information

It is not easy to process the text from social media as they contain very little contextual information and assume too much implicit knowledge.

2.1.11 Non availability of resources

Like annotated data or parser, pos taggers etc.

3. Related Work

Social media text normalization is a challenging problem and attracted the attention of researcher from all over the globe. Two approaches in literature are found to handle text in social media. Either transform the input text (i.e. normalize it) or transform the tools. While the second option is feasible mostly for statistically trained tools, the first option should work for any tool, statistical and rule-based. Both statistical based and rule based techniques are found in literature for handling un-normalised text.

For normalization of Chinese social media text, [1]investigated informal phrase detection, and[2] mined informal-formal phrase pairs from Web corpora[3], have produced a system for normalizing Short Message Service mobile phone texts, which share many of the characteristics of the casual English, such as non-standard short-forms of words, creative phonetic or stylistic spelling, and punctuation omission. They proposed a noisy channel model consisting of different operations: substitution of non-standard acronyms, deletion of flavor words, and insertion of auxiliary verbs and subject pronouns. Supervised approach was used by[4] to perform text normalization of informally written email messages using CRF. HMM modle was used by[5] to perform word-level normalization[6]. In their work for the CAW 2.0 project introduced an approach using a n - gram based SMT system and were able to produce syntactically correct sentences from input with a high frequency of misspelled words and Internet slang[7]. Used an unsupervised noisy channel model considering different word formation processes[8]. Proposed Casual English Conversion System (CECS) which exploit combination of automated and manual techniques[9]. Uses a classifier to detect ill-formed words, and generates correction candidates based on morphophonemic similarity[10]. Normalized social media texts incorporating orthographic, phonetic, contextual, and acronym factors[11]. Only dealt with SMS abbreviations[12]. Present an approach to text normalization that uses a language model-based automatic correction selector, built on top of a pre-existing spellchecker[13]. Designed a system combining different human perspectives to perform word-level normalization[14]. Describe a method for automatically constructing a normalization dictionary that supports normalization of micro-blog text through direct substitution of lexical variants with their standard forms[15]. Proposed unsupervised learning of the normalization equivalences from unlabeled text. The proposed approach uses Random Walks on a contextual similarity bipartite graph constructed from n-gram sequences on large unlabeled text corpus[16]. Target out-of-vocabulary words in short text messages and proposed a method for identifying and normalizing lexical variants based on morphological and phonetic variation[17]. Normalized Spanish SMS messages using a normalization and a phonetic dictionary[18]. Present a unified unsupervised statistical model for text normalization[19]. Argue that along with word replacement other normalization operations should also be performed, e.g., missing word recovery and punctuation correction. They propose a novel beam-search decoder for normalization of social media text for MT[20]. Proposed a method for learning normalization rules from machine translations of a parallel corpus of microblog messages. It is data driven approach.

4. Methodology

Normalizing Social media text in Gurumukhi but written in roman script involves transliteration and then conversion to canonical form. Better the accuracy of transliteration, better the result of normalization. We experimented with a rule based roman to Gurmukhi transliteration system and treated it as our baseline. As we will show in the experiments in Section (5), Rule-based transliteration method proved to be inadequate for social media domain normalization for many reasons as cited in section 2. In this section, baseline method is discussed first followed by a new algorithm for improving the results of normalization.

4.1 Rule-Based Transliteration

When Punjabi is written with Roman script, sometime two character of roman alphabets are required to denote some of the characters in Gurmukhi e.g. kh (ਖ), gh (ਘ), ch (ਚ) etc. Input text is preprocessed to replace all such occurrences with their counterpart in Gurumukhi script. A mapping table (Table 1.) is used for mapping a given roman character to Gurumukhi alphabet.If input text consists of single letter then it is mapped according to first column otherwise mapped according to second column. It will cover all the abbreviation like B J P "ਬੀਜੇਪੀ". It also covers word shortening cases like use of letter v for "ਵੀ".It is possible to have multiple characters for a given roman alphabet, but most frequently used one character is selected in this mapping. The converted string may include some character sequences that are invalid in Gurumukhi script. These invalid sequences are corrected using rules like if a character appears consecutively, then first occurrence is replaced by diacritic mark "ੱ" (Adhak) which denotes stress on second character. Similarly if a word starts with diacritic mark, then replace diacritic with its full vowel.

Table 1. Roman to Gurumukhi Character Mapping

Roman Character	Replaced By		Roman Character	Replaced By		Roman Character	Replaced By	
	if single character	Otherwise		if single character	Otherwise		if single character	Otherwise
a	ਆ	ਾ	j	ਜੇ	ਜ	s	ਐਸ	ਸ
b	ਬੀ	ਬ	k	ਕੇ	ਕ	t	ਟੀ	ਤ
c	ਸੀ	ਕ	l	ਐਲ	ਲ	u	ਯੂ	ੁ
d	ਡੀ	ਦ	m	ਐਮ	ਮ	v	ਵੀ	ਵ
e	ਈ	ੇ	n	ਐਨ	ਨ	w	ਡਬਲਯੂ	ਵ
f	ਐਫ	ਫ	o	ਓ	ੋ	x	ਐਕਸ	ਕ
g	ਜੀ	ਗ	p	ਪੀ	ਪ	y	ਵਾਈ	ਯ
h	ਐਚ	ਹ	q	ਕਯੂ	ਕ	z	ਜ਼ੈਡ	ਜ਼
i	ਆਈ	ਿ	r	ਆਰ	ਰ			

4.2 Proposed Algorithm

Simple rule based technique is not sufficient for converting text from roman to Gurumukhi script due to the issues discussed in section 2. This paper discuss the handling of first three cases viz. different ways of writing, phonetic similarity and text shortening.A new algorithm is proposed to handle these cases. This algorithm require following resources:

- List of Punjabi Words.
- n-grams: Bigrams and trigram model has been used.
- List of English words written in Gurumukhi script. This list is required because the word may belong to English language.
- Roman character to Gurumukhi character mapping table (Table 2). This table is different form Table 1 in the sense that it maps all the possiblecharacter in target script for a given character.

Table 2. Roman to Gurumukhi Multiple Character Mapping

Roman Character	Mapped character	Roman Character	Mapped character	Roman Character	Mapped character
a	~, ਅ, ਆ, ਾ	j	ਜ	s	ਸ
b	ਬ	k	ਕ	t	ਤ, ਟ
c	ਕ, ਚ, ਸ	l	ਲ	u	ਉ, ਊ, ੁ, ੂ
d	ਦ, ਡ	m	ਮ, ੰ,ਂ	v	ਵ
e	ੲ,ਇ, ਈ,ਏ, ਐ, ੈ ,ੇ, ਿ ੀ	n	ਨ, ੰ,ਂ	w	ਵ
f	ਫ	o	ੳ, ਓ, ਅ, ੋ, ੌ	x	--
g	ਗ, ਜ	p	ਪ	y	ਏ, ੲ,
h	ਹ	q	ਕ	z	ਜ, ਜ਼
i	ੲ,ਇ, ਈ,ਏ, ਿ, ੀ	r	ਰ, ੜ		

The algorithm generates all the combination of each word. General tree has been used for this purpose. As one character in roman may map to multiple characters in Gurumukhi, different branches can be stemmed out of a given node. Every branch in this tree leads to new combination. Some of the combinations are not valid like the combination starting from diacritic marks. For example one of the combinations for "ik" is "ਿਕ" which is invalid in Punjabi. Such combinations are pruned from the list of all combinations. To handle the text shortening, resulting list is used to generate new combinations. It is observed that in roman script, people generally write shortened form e.g. "bht" for "ਬਹੁਤ". Above algorithm will generate the word "ਬਹਤ". To generate "ਬਹੁਤ" from "ਬਹਤ", a coding scheme is applied. Every Gurumukhi character except diacritic symbols has been given a code. Thus ਬਹੁਤ and ਬਹਤ gets same code. All words from Punjabi word list, having same code are selected and added to the candidate list. Finally, if word also appears in

English wordlist then corresponding Punjabi word from database is also added to the candidate list.

After generating all the possible combinations of a word that word is selected whose probability of occurrence in given context is high. Bigram model is used for this purpose. Besides we also select those words which are generated through English wordlist. The given word can be a name of person or place. So its default mapping to Gurumukhi script has also been included. Again a general tree has been used for generating all possible sentences from candidate words. Depth first traversing this tree generates a list of possible sentences. HMM model is used to select one sentence out of the generated possibilities.

5. Experimental Setup

Data from public facebook pages was collected. A test set of 1000 sentences was constructed. This test set had been corrected by a native human annotator. The human annotator normalizes noisy words to its canonical form in a consistent way according to the evidences in the context. This test set is passed through baseline system. Then it is passed through new system. The precision, recall and F-measure are as shown below:

Table 3. Evaluation Result

System	Precision	Recall	F-measure
Baseline	34.7	16.1	21.9
Proposed	52.3	25.9	34.6

6. Result and Discussion

Table 4. Some Sample Output

Input: Sher put ikbahutaa Baseline: ਸ਼ੇਰਪੁਤਕਿਬਹੁਤਾ Proposed Algorithm: ਸ਼ੇਰਪੁੱਤਇਕਬਹੁਤਐ
Input: Angrezich gall krdi Punjabi kudi Baseline: ਐਨਗਰੇਜ਼ਚਿਗੱਲਕਰਦਪੁਿਨਜਬਕੁਿਦਾ Proposed Algorithm: ਅੰਗਰੇਜ਼ੀਚਗੱਲਕਰਦਪਿੰਜਾਬੀਕੁੜੀ
Input: Wahbaibeant sea siraalaa to Baseline: ਵਹਬੈਬੀਨਤਸੀਸਰਿਾਲਾਤਾ Proposed Algorithm: ਵਾਹਬਈਬੀਨਤਸੀਸਰਿਾਲਾਤਾ
Input: Kithe raja bhojtekithegangutelii Baseline: ਕਥਿਰਜਾਭੋਜਤੇਕਥਿਗਨਗੁਤੇਲਾਂ Proposed Algorithm: ਕਥਿਰਾਜਾਭੋਜਤੇਕਥਿਗੰਗੁਤੇਲੀ

The results show that baseline approach is not fit for social media text. The reason behind is that baseline approach is unable to capture the informal way of writing. The proposed algorithm impacts the performance of system positively. The improved F-score implied that the proposed algorithm has some success in capturing the variations in social media text. Some of the outputs are shown in Table 4.

7. Future Scope

We introduced a social media text normalization system for Punjabi Language and attempted to normalize the Punjabi text written in roman script to Gurumukhi script. We tried to solve some of the issues of normalizing text and proposed an algorithm. The algorithm improves the output but still lacks in various aspects. In future, we will concentrate on other issues and try to develop a system for normalizing Punjabi Text.

8. References

1. Xia Y, Wong KF, Gao W. NIL is not nothing: Recognition of Chinese network. Proceedings of 4th SIGHAN Workshop on Chinese Language Processing; 2005.
2. Li Z, Yaorwsky D. Mining and modeling relations between formal and informal Chinese phrases from web corpora. Proceedings of EMNLP; 2008.
3. Aw A, Zhang M, Xiao J, Su J. A phrase-based statistical model for SMS text normalization. Proceedings of COLING/ACL on Main Conference Poster Sessions (COLING-ACL '06), Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. p. 33–40.
4. Zhu C, Tang J, Li H, Ng HT, Zhao TJ. A unified tagging approach to text normalization. Proceedings of ACL; 2007.
5. Choudhury M, Saraf R, Jain V, Mukherjee A, Sarkar S, Basu A. Investigation and modeling of the structure of texting language. International Journal on Document Analysis and Recognition. 2007; 10(3):157–74.
6. Henriquez C, Hernandez A. Angram-based statistical machine translation approach for text normalization on chat-speak style communications. Proceedings of CAW2. 0, Madrid, Spain; 2009. p. 1–5.
7. Cook P, Stevenson S. An unsupervised model for text message normalization. Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity, Boulder, Colorado: ACL; 2009. p. 71–8.
8. Clark E, Araki K. Text normalization in social media: Progress, problems and applications for a pre-processing system of casual english. Procedia - Social and Behavioral Sciences. 2011; 27:2–11.
9. Han B, Baldwin T. Lexical normalisation of short text messages: makn Sens a #twitter. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon: Association for Computational Linguistics; 2011. p. 368–78.

10. Xue Z, Yin D, Davison BD. Normalizing microtext. Proceedings of the AAAI Workshop on Analyzing Microtext; 2011.
11. Pennell DL, Liu Y. A character-level machine translation approach for normalization of SMS abbreviation. Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand; 2011.
12. Melero M, Costa-Jussa MR, Domingo J, Marquina M, Quixal M. Holaaa! writin like u talk is kewl but kinda hard 4 NLP. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA); 2012.
13. Liu F, Weng F, Wang B, Liu Y. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon: ACL; 2011. p. 71–6.
14. Han B, Cook P, Baldwin T. Automatically constructing a normalisation dictionary for microblogs. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea: Association for Computational Linguistics; 2012;. p. 421–32.
15. Hassan H, Menezes A. Social text normalization using contextual graph random walks. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia Bulgaria: ACL; 2013. p. 1577–86.
16. Han B, Cook P, Baldwin T. Lexical normalization for social media text. ACM Transactions on Intelligent Systems and Technology (TIST). 2013; 4(1):1–27.
17. Olivia J, Serrano JI, Castillo MDD, Igesias A. A SMS normalization system integrating multiple grammatical resources. Natural Language Engineering. 2013; 19(1):121–41.
18. Yang Y, Eisenstein J. A log linear model for unsupervised text normalization. Proceedings of Empirical Methods in Natural Language Processing, ACL; 2013. p. 61–72.
19. Wang P, Ng HT. A beam search decoder for normalization of social media text with application to machine translation. Proceedings of NAACL-HTL, Atlanta, Georgia; 2013. p. 471–81.
20. Ling W, Dyer C, Black AW, Trancoso I. Paraphrasing 4 microblog normalization. Proceedings of Empirical Methods in Natural Language Processing, ACL; 2013. p. 73–84.