2012 IEEE 24th International Conference on Tools with Artificial Intelligence

A Rule-based Model for Normalization of SMS Text

Osama A. Khan and Asim Karim
Department of Computer Science, SBASSE
Lahore University of Management Sciences (LUMS)
Lahore, Pakistan
Email:{oakhan, akarim}@lums.edu.pk

***Abstract*—SMS are short-length text documents written in a colloquial style. SMS text processing is challenging because of low signal-to-noise ratio and multi-varied text composition in terms of language, vocabulary, style and quality. These challenges can be overcome by robust text normalization, which is a necessary step before any technique can be applied and evaluated on such data. In this paper, we present a rule-based model for multi-lingual SMS text normalization focusing on messages written in Romanized Urdu and English. Urdu in contrast to English is a morphologically rich language (MRL), i.e. it produces a very large number of word forms for a given root form, while Romanized Urdu is a way of writing Urdu in Latin script which does not follow standard rules for systematic communication. Hence, normalization or standardization of multi-lingual SMS text offers challenges associated with SMS text, multi-lingualism, MRLs and Latin script. Our SMS standardizer is based upon a tuned set of rules that range over various domains of natural language processing, and which tackle the challenges mentioned above effectively. We then implement the standardizer in the application of Keyword Extraction from SMS messages, where it produces significant improvement in performance by upto 23% in F-measure.**

***Index Terms*—Text Normalization, SMS, Rule-based model, Romanized Urdu, Keyword Extraction**

I. Introduction

Short Message Service (SMS) texts on cellular networks are short-length textual documents exchanged among users on the Web and/or their mobile devices. SMS-based services and social networks have become very popular for rapid information sharing and communication in recent years. Such textual documents have a short length of upto 160 characters, low signal-to-noise ratio and a multi-varied composition. The last characteristic has several dimensions: language, vocabulary, style and quality, which comprise multi-linguality, abbreviations, contractions, acronyms, proper nouns, multi-word names, compound words, slangs, colloquialisms, multi-styled words and typos. It is therefore necessary to propose and evaluate novel text normalization techniques for this noisy and varied document type. We build a rule-based model that handles noise and multi-varied composition in SMS text effectively as discussed in detail in Section III. Table I displays classes and examples of multi-varied composite features of SMS text. Throughout the paper we supplement Romanized Urdu words with corresponding English translations in brackets [].

TABLE I: EXAMPLES OF VARIOUS FEATURES OF SMS TEXT

Feature	Class	Example
Noise	**Noise**	βetter :) gr8
Multi-linguality	**Language**	Friday hay [It is Friday]
Abbreviation	**Vocabulary**	Interpol => International Police
Contraction		tc => take_care
Acronym		CNN => Cable_News_Network
Proper Noun		Paris
Multi-word Name		Warren Buffet => Warren_Buffet
Compound Word		chit chat => chit_chat
Slang	**Style**	lol => laughing_out_loud
Colloquialism		wanna => want_to
Multi-styled Word		ok, oke, okey, okdk => okay
Typo	**Quality**	acheive => achieve

Multi-lingualism in SMS text produces messages written in multiple languages and scripts, with combinations of morphologically and non-morphologically rich languages used for texting. Morphologically rich languages (MRLs) are different from non-MRLs such that the former are characterized by highly productive morphological processes (e.g. inflection, agglutination and compounding) which result in a large number of word forms for a given root form, and where grammatical relations are independent of word positions inside a document, while in non-MRLs grammatical relations are defined positionally (e.g. in English). MRLs are often resource-scarce while sophisticated processing tools are usually available for non-MRLs. For normalization of such documents, either language-independent tools are employed or else separate models for each language are required. Also, due to multi-varied nature of microblogs external resources like Wikipedia and WordNet are not applicable. Our rule-based model meets the challenges associated with both multi-lingual text and MRLs as discussed in detail in Section III.

Urdu, the national language of Pakistan and official language of five states of India, is one such MRL which is derived together from three other MRLs, namely Arabic, Persian and Turkish. Romanized Urdu is one style of writing Urdu in Latin script, and it is independent of

1082-3409/12 $26.00 © 2012 IEEE
DOI 10.1109/ICTAI.2012.91

various rules that a language is constructed upon. It is mostly used in colloquial styled conversations via chat, microblogs or SMS, due to informal characteristics of both Romanized Urdu and social media, and also due to the absence of Calligraphic Nastaliq [1] feature in a majority of these types of services. Hence, normalization of SMS text needs to deal with a variety of challenges related to SMS text, multi-lingualism, MRLs and Latin script. Designing an automated system for such a process without human supervision seems to be a complicated task, due to which we develop a manual standardizer that tackles all such challenges as discussed in detail in Section III. The following is an SMS message typed with a mix of Romanized Urdu and English, sent over a social network:

> *aj friday hay is jaldi chutti hogayi, aur wesay main mob lekare jata hun. tm ne kal program dekha tha kya?*
> [It is Friday today, therefore I got off early, otherwise I take mobile with me. Did you see the program yesterday?]

Normalization of SMS text consists of two sub-problems, namely developing a repository containing all non-standardized variations of words that are present in a limited sized document collection, alongwith their unique standardized forms, and then updating the repository by mapping a new non-standard word (NSW) detected in a processed document to its standardized form already present in the repository. In this paper, we propose a novel solution to the first sub-problem, which builds a repository from a large multi-lingual (Romanized Urdu and English) SMS collection by utilizing a sophisticated rule-based model that is developed using various features of both natural language processing and SMS text.

The rule-based SMS standardizer for multi-lingual text constructs a repository by utilizing a data collection that comprises over 260,000 SMS messages dated between October 9th, 2008 and December 12th, 2008, and acquired from a Pakistani SMS-based online social network. Due to privacy reasons we do not disclose the name of this network. The SMS messages contain a total of 120,000 words out of which 51% are found to be NSWs, thus verifying the significance of normalization of SMS text. The built up repository contains 61,069 NSWs that are mapped to 22,777 unique forms. On average, about 3 non-standardized variants are mapped to a unique standard form, with a maximum of 79 variants existing for the word 'mahabbat' [love]. Also, about 53% unique words possess a single variant, 17% words have two variants, 10% words are linked with three variants, while only 20% words are associated with four or more non-standardized variants.

One of the applications for which text normalization is considered a prerequisite is Keyword Extraction, which is the process of extracting terms as keywords from a document. In this paper, we utilize term frequency, inverse document frequency (TFIDF) that is considered a baseline for performance evaluation of keyword extraction from microblogs [2], to measure improvement in the performance when the SMS standardizer is incorporated as a preprocessing procedure. We also study the relationship between noise reduction through SMS standardizer and vocabulary size of the collection.

The rest of the paper is organized as follows. In Section II we provide literature review of the work already carried out in the area of Microblog Normalization. We present our rule-based SMS standardizer for multi-lingual text in Section III. In Section IV we implement the SMS standardizer in the application of Keyword Extraction. We make concluding remarks and outline key future directions in Section V.

II. Related Work

The first solution for normalization of microblogs was devised by Aw et al. [3] in the form of a framework based on Noisy Channel Model (NCM) and Machine Translation (MT). They embedded Expectation Maximization, Viterbi Search, Edit Distance, SMS lingo dictionary and SRILM toolkit in their normalization framework, tested it on an SMS corpus, and then evaluated it as a prerequisite in the application of machine translation from English (a non-MRL) to Chinese (an MRL) SMS messages. Later on, Kobus et al. [4] presented three different techniques for normalization of French (an MRL) SMS messages; first using MT, the second utilizing Automatic Speech Recognition, and third being a combination of these two techniques. In the recent past, Pennell and Liu [5] devised a rule-based, Maximum Entropy and Maxent algorithm based technique for transformation of NSWs in microblogs to their standard forms and vice versa, and implemented it on SMS messages and Twitter tweets. They later formulated a CRFs based technique for this dual task and evaluated it on a Twitter corpus [6]. Lately, Beaufort et al. [7] presented another technique for normalization of French SMS messages, which incorporated rule-based model, MT, and Finite-state Machines and used a lexicon. More recently, Liu et al. [8] devised an NCM and CRFs based technique for microblog normalization by embedding Letter-Phoneme Alignment algorithm and two dictionaries. Lately, Han and Baldwin [9] introduced a joint model based on Double Metaphone Algorithm and Support Vector Machines for the dual task of identification and normalization of NSWs in microblogs by using NYT English corpus and two dictionaries, and evaluated their developed technique on SMS and Twitter corpora.

III. Rule-based SMS Standardizer for Multi-lingual Text

In this section, we discuss in detail solution to the first sub-problem of normalization of SMS text (see Section I) in the form of a rule-based SMS standardizer. A typical rule-based system consists of three major components,

namely a rule base, an inference engine and a user interface, using which the system stores and interprets knowledge in order to generate useful decisions. The rule base consists of a list of rules and facts which alongwith system input are utilized by the inference engine to make decisions. The user interface provides communication between user and the system, e.g. to resolve conflicts resulting from interactions between system input and the rule base. An automated rule-based system utilizes rules defined in the form of logical propositions with antecedent and consequent components. However, in our case significant challenges from multiple domains make it very difficult to develop a fully automated rule-based system that uses predefined propositional rules for text normalization. Therefore, we build a semi-autonomous rule-based model that utilizes a rule base, an inference engine as well as human interaction in order to normalize multi-lingual SMS text effectively.

The SMS collection that we utilize in this work contains messages written predominantly in Latin script, but in addition to English it contains significant fractions of messages written in Romanized Urdu and local languages mixed with English. In a manual inspection of 2,000 messages we observe that about 49% and 26% of messages are written primarily in English and in Romanized Urdu respectively, another 20% of messages utilize a mixture of Romanized Urdu and English, while the remaining 5% of messages employ one of the three local languages. Our rule-based standardizer is developed specially for normalization of words typed in Romanized Urdu and English, while leaving the words in other languages unstandardized in the SMS collection. The SMS collection is first tokenized via space delimitation and a list of unique vocabulary words is generated, which is then transformed into lower case in order to reduce the vocabulary size. Normalization procedures are implemented sequentially on the vocabulary list, thus building a repository out of it which contains all original NSWs in the vocabulary list alongwith their associated standardized forms.

Here, we first define the challenges of SMS text, multilingualism, MRLs and Latin script and then provide solutions to these challenges in the form of normalization procedures which are based on interactions between human expert, rule base and inference engine in the following subsections. The challenges are listed in the order that they are implemented, and the rules (from the rule base) employed for their solutions are listed in the order that they are selected to be fired either by the human expert or inference engine, in order to avoid a conflicting set of rules. Figure 1 depicts the architecture of the rule-based SMS standardizer that handles these challenges sequentially. The rules are designed while taking into consideration the complexities associated with these challenges, and thus provide near optimal transformations of NSWs into their standardized forms. All rules are assumed to be automated and handled by the inference engine, except

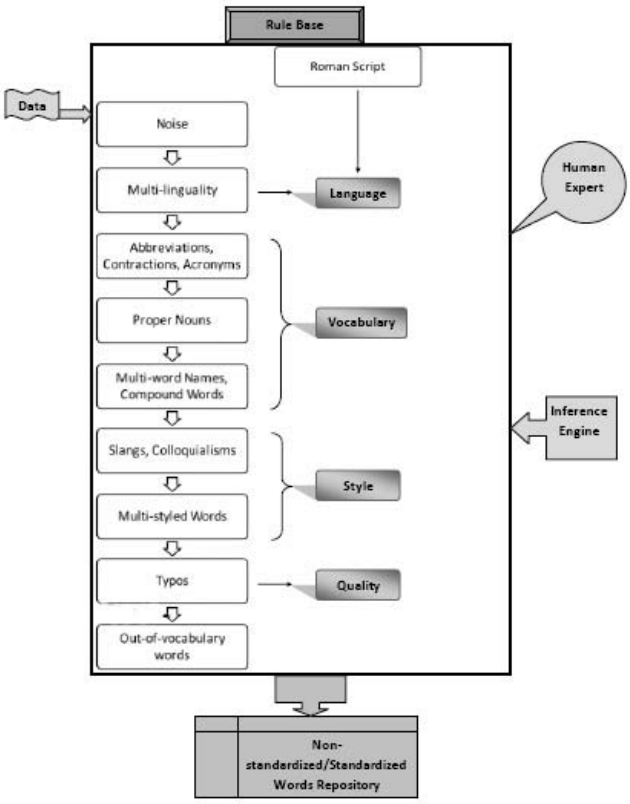

Fig. 1: Architecture of Rule-based SMS Standardizer

for those where human supervision is explicitly stated. Both the human expert and inference engine consult a unique list associated with each rule mentioned in the following subsections where required, which contains samples of words drawn by human expert from over the entire SMS collection (see Section I). For example, for rule 1 of Section III-A the inference engine maps all symbols present in text to English letters by looking up a list that contains pairs of symbols and their associated letters.

A. Noise

Noise is defined by the presence of any of the following in text: Symbols ∼ @ # $ % ^ & * + = { } [ ] \ | / < >, Punctuation . , : ; ’ ” ( ) ! ? -, or Numerals 0, 1, 2, …, 9.

1) Any symbols (non-alphabetic characters) present in the vocabulary words are transformed to appropriate letters in English alphabet, like ‘β’ is replaced by ‘b’, ‘è’ by ‘e’, ‘£’ by ‘l’, ‘ï’ by ‘i’, and ‘§’ by ‘s’. Symbol ‘_’ is the only character that is allowed in standardized words.

2) Mis-encoded symbol ‘ï¿1/2’ from UTF-8 text format is replaced by any letter in the English alphabet which best forms the word by its replacement, using human intuition. If no appropriate word is formed by replacement of ‘ï¿1/2’ with any English letter, then letter ‘i’ replaces it in the word.

3) If a word contains the symbol 'ï¿1/2' more than once in it, then human intuition is first used to derive the most frequently used English word from it by replacing its symbols with (possibly) different letters. If intuition doesn't yield any appropriate English word, then either the symbols are replaced by letter 'i' if the word is too long to figure out a valid replacement from it, or in case the word is short symbols are replaced by letters by the human expert who tries each letter in English alphabet in that order, to judge an appropriate replacement.

4) All punctuation characters and numerals present in words are removed.

B. Romanized Urdu vs. English

The human expert identifies Romanized Urdu and English vocabulary words and employs the rules mentioned in this subsection for preliminary standardization of words belonging to each language accordingly. For rules 1 and 2, the next source is consulted only if the previous source does not result in any appropriate replacement.

1) American English is used for standardization of all English words by consulting the following sources in the given order.

a) WordWeb[1], software that has built-in English thesaurus and dictionary.

b) WIKIPEDIA[2], multi-lingual web-based and free-content encyclopedia.

c) Google[3], multi-lingual web search engine.

2) All Romanized Urdu words are found an appropriate replacement in Urdu by consulting the following sources in the given order.

a) Urdu English Dictionary[4]

b) iJunoon[5], web portal that performs Romanized Urdu to Urdu transliteration

c) HAMARI WEB[6], web portal that provides translation services from several languages to Urdu

d) urbaN DICTIONARY[7], multi-lingual dictionary that contains slangs and colloquial words

e) Google[3]

f) WIKIPEDIA[2]

g) Human expert, who assigns the best replacements for confusing words according to his judgment

The substituted Urdu words are then transliterated into standardized Romanized Urdu using the rules defined ahead in this subsection.

3) Table II provides a mapping for each letter in Urdu alphabet to its corresponding English alphabet letter(s).

4) Table III describes the mapping of some letters in Urdu alphabet to their corresponding letters in English alphabet, when the former are used in the middle or end of Urdu words.

TABLE II: URDU-ENGLISH ALPHABET TRANSCRIPTIONS

S#	Urdu alphabet	English alphabet	S#	Urdu alphabet	English alphabet
1	ا	a, aa	21	ص	s
2	آ	aa	22	ض	z
3	ب	b	23	ط	t
4	پ	p	24	ظ	z
5	ت	t	25	ع	a, u
6	ٹ	tt	26	غ	gh
7	ث	s	27	ف	f
8	ج	j	28	ق	q
9	چ	ch	29	ک	c, k
10	ح	h	30	گ	g
11	خ	kh	31	ل	l
12	د	d	32	م	m
13	ڈ	dd	33	ن	n
14	ذ	z	34	ں	n
15	ر	r	35	و	w
16	ڑ	rr	36	ء	a
17	ز	z	37	ہ	h
18	ژ	zsh	38	ھ	h
19	س	s	39	ی	y
20	ش	sh	40	ے	y

5) Sign 'ّ' (tashdeed) when used over a letter in the middle of an Urdu word results in double usage of its corresponding letter(s) in English alphabet, as in 'ٹھڈّا' 'thuddda' [kick] and 'بچّہ' 'bachchah' [child].

6) Sign 'ّ' (tashdeed) when used over a letter at the end of an Urdu word results in single usage of its corresponding letter(s) in English alphabet, as in 'اہمّ' 'aham' [ahem] instead of 'ahamm'.

7) Sign 'ٰ' (double zabar) when used over a letter at the end of an Urdu word results in single usage of its corresponding letter(s) in English alphabet, as in 'اندازاً' 'andaazan' [approximately] instead of 'andaazann'.

8) Plurals of Romanized Urdu words that end in 'i' are formed by affixing 'yaan' or 'yon' to those words without replacing 'i' with 'ee', as in 'achchhaai' [quality] to 'achchhaaiyon' [qualities] or to 'achchhaaiyaan' [qualities].

9) Some words seem to have equal chances of getting any of the considered replacements either in the same language or in different languages. In this case, the word is left intact, like 'ie' can be transformed either to 'آئے' 'aaey' [came] or to 'internetexplorer' with almost equal probability.

C. Abbreviations, Contractions and Acronyms

An abbreviation is a shortened form of a word or phrase, a contraction is an abbreviation which is shortened by omission of internal letters of a word or phrase, while an acronym is an abbreviation which is formed from the initial components of a word or phrase (see Table I for examples).

1) All abbreviations and their subclasses are expanded to their full forms, e.g. 'dnt' to 'donot', 'km' to 'kilometer', etc.

[1] http://wordweb.info/
[2] http://en.wikipedia.org/
[3] http://google.com
[4] Ferozsons (Pvt.) Ltd. (1998)
[5] http://ijunoon.com
[6] http://hamariweb.com
[7] http://www.urbandictionary.com/

TABLE III: URDU-ENGLISH LETTER PAIRS FOR MIDDLE AND END POSITIONS OF WORDS

Urdu letters	Middle of word	Example	End of word	Example
ا	aa	مکان makaan [house]	a	بتانا bataana [to tell]
ی (zer)	ee, i, ai, ae	آمین aameen [Amen], اچھائیوں achchhaaiyon [qualities], کریں karain [to do], بیٹھنا baetthna [to sit]	i	آخری aakhri [last]
(zer)	i	عامل aamil [adminstrator]	-	-
ع (zer)	i	موقع دے mauqiday [give opportunity]	i	شائع shaai [publish]
و	o, oo, au	سوچ soch [thinking], ابّوجی abbooji [daddy], عورت aurat [woman], اکلوتا iklauta [sole]	o, u	بولو bolo [speak], آبرو aabru [honor]
ؤ	o, oo	آؤگے aaogay [will come], آؤنگا aaoonga [will come]	o	بتاؤ bataao [tell]
(pesh)	u	دشمن dushman [enemy]	-	-
ئی	ei	جائیگا jaaeiga [will go]	i	بتائی bataai [told]
ے	-	-	ay	بھرے bharay [filled up]
ئے	-	-	ey	بتائے bataaey [told]
اے (sound)	ae	احترام aehtiraam [respect], اہل aehl [capable], اعتبار aetibaar [trust]	-	-
ء	aa	اونچائی oonchaai [height]	*silent*	انبیاء ambiya [prophets]
ة	-	-	h	آیۃ aayah [verses]

2) An abbreviation or its subclass when existing as part of a word is expanded to its full form while the remaining word is kept intact, e.g. 'idnt' to 'idonot'.

3) An abbreviation or its subclass when affixed with 'ian' at the end is not expanded to its full form, as in 'nedian' and 'fastian'. These words are usually ascribed to people belonging to academic institutions whose abbreviated forms are used here.

D. Proper Nouns

A proper noun is a noun representing a unique entity, where a noun is a word representing such entities in general, e.g. 'John' (person), 'Berlin' (place), 'Microsoft' (organization), 'Lion' (animal), etc. Here, all proper nouns are written in lowercase as are other words in text.

1) All proper nouns except names of people are standardized in English as per rule 1 of Section III-B.

2) Names of people in Romanized Urdu are standardized as per rules 2 - 7 of Section III-B. If all sources mentioned in rule 2 of Section III-B are exhausted for a person's name, then sources mentioned in rule 1 of the same subsection are consulted for name standardization. If sources mentioned in this rule are also exhausted, then the name is left as it is.

3) Names of people in English are standardized as per rule 1 of Section III-B. If all sources according to this rule are exhausted for a person's name, then the name is left as it is.

E. Multi-word Names and Compound Words

A compound word is formed by joining two or more words together such that a new meaning emerges that is quite different from the meanings of the words in isolation, e.g. underworld. Multi-word proper nouns or names are one class of compound words.

1) Punctuation character hyphen '-' is replaced by underscore '_' in all compound words.

2) When two or more words are written together in a compound word without any joining characters between them, they are standardized according to the rules governing their respective languages as mentioned in Section III-B, without inserting any '-' or '_' between them, as in 'assalaamoalaekum' [hello].

3) When two or more words are written together in a compound word with joining character(s) between them, the words are standardized according to the rules governing their respective languages as mentioned in Section III-B, while the joining characters are replaced by a single '_' in the same positions as in the original phrase, like 'asalaam-o-alaikum' to 'assalaam_o_alaekum' [hello], and 'aachanuk_____ma' to 'achaanak_main' [surprisingly].

4) When a joining character exists between two words in a compound word, e.g. 'keeey_gaey' [carried out], the words are considered disjoint while applying standardization rules mentioned in Section III-B upon them. However, in case of a compound word without a joining character between two words, e.g. 'keeeigaey' [carried out], although the standardization rules are applied to both the words separately they are still considered joint words. Therefore the two compound words mentioned above although being variations of the same word are spelled differently after their respective transformations as per rule 4 of Section III-B.

5) Consecutively repeated '_' in compound words are considered blanks and are therefore tried at best to be filled up with appropriate letters or words by the human expert, as from 'i_____u' to 'iloveyou'.

6) Romanized Urdu compound words are usually written with joining characters 'e', 'i' and 'o' besides using '_', as in 'aaghaaz_e_mohabbat' [beginning of love], 'nawa_i_waqt' [voice of time] and 'roz_o_shab' [day and night] respectively. Such words when are written incomplete are transformed accordingly by the human expert, as from 'rasmo' to 'rasm_o_' [tradition of].

7) Romanized Urdu compound words affixed with 'bay' [without] in the beginning usually refer to antonyms. Exceptions are given to such compound words in following rule 4 of Section III-B, and the word 'bay' is spelled in

its original form instead of converting it into 'bai', as in 'baywafa' [unfaithful] rather than 'baiwafa'.

8) Single words that possess a joining character(s) at their ends get transformed according to standardization rules mentioned in Section III-B while the joining character(s) gets removed, like 'aadee_' to 'aadi' [accustomed] and 'l___' to 'l'.

9) Single words that seem to have been deliberately compounded get their joining characters removed by the human expert, like 'l_i_p' to 'lip'.

10) Some single words include '_' in between their letters in such a way that the user intention is to get the '_' filled up with an appropriate letter in order to form a valid word. Such words are standardized by the human expert by first forming valid words from them, like 'b___ddh___' to 'buddhu' [fool], and then by applying rules mentioned in Section III-B.

F. Slangs and Colloquialisms

A colloquialism is a word or phrase that is employed in conversational or informal language, while a slang is a colloquialism which is not considered standard in a language but is acceptable when used socially (see Table I for examples).

1) All slangs including vulgar words are standardized according to rules associated with the language that they belong to as mentioned in Section III-B, and are considered part of the vocabularies of their respective languages.

2) Colloquialisms used for family relationships in English are transformed to their respective formal words, like 'ma', 'mama' and 'mommy' get all transformed to 'mother'.

3) Colloquialisms used for family relationships in Romanized Urdu are kept intact, like 'abbu' [daddy] and 'abba' [daddy] do not get transformed to 'baap' [father].

G. Interjections

An interjection or exclamation is a word or phrase used to express an emotion, sentiment or a pause in a sentence, e.g. 'oh dear', 'sorry', 'um', etc.

1) The following are considered as standard interjections and are therefore left intact:

a) Romanized Urdu expressions like 'aachhi' [sneezing sound], 'aachhu' [sneezing sound], 'aaha' [great], 'abay' [hey], 'chhi' [shit], 'oho' [oh], 'thu' [shit], 'uf' [oh], etc.

b) English expressions like 'ahem', 'alas', 'aw', 'bubye', 'cheers', 'er', 'haha', 'hey', 'hurrah', 'oh', 'ugh', 'wow', etc.

2) The following transformations take place to standardize interjections:

a) Romanized Urdu expressions like 'tata' to 'bye'.

b) English expressions like 'ok', 'oka', 'okdk', 'oke' and 'okey' to 'okay', 'shhhhshh', 'shish' and 'sss' to 'shh', 'yea', 'yeah', 'yep' and 'yup' to 'yes', etc.

H. Seasons, Months and Days

1) The following transformations take place to standardize seasons, months and days in a year.

a) Seasons 'aut' to 'autumn' and 'spr' to 'spring'.

b) Months 'feb' to 'february', 'apr' to 'april', 'jun' to 'june' upto 'dec' to 'december'.

c) Days 'mon' to 'monday', 'tue' to 'tuesday', and 'fri' to 'friday'.

2) Abbreviations of the rest of times of year correspond to either standard English or Romanized Urdu words, and are therefore not expanded to their full forms but are instead treated as per rules mentioned in Section III-B.

I. Latin Numerals

Latin numerals use combinations of seven letters from the Latin alphabet to signify numerical values (i=1, v=5, x=10, l=50, c=100, d=500, m=1,000).

1) Latin numerals ('one' to 'nineteen') are transformed to their English names only when they are the only content present in words, e.g. 'ii' to 'two', 'x' to 'ten' upto 'xix' to 'nineteen'. Numerals 'i' and 'v' are made exceptions to this rule, where the former is left intact and the latter is transformed to 'we' instead of 'five'.

2) When Latin numerals ('one' to 'nineteen') are part of words that contain other content too, then these words are transformed according to standardization rules mentioned in Section III-B, as from 'ii_' and 'i_i' to 'ii', which is not further transformed to English name of the Latin numeral ('two').

3) Latin numerals 'xx' (twenty), 'xxx' (thirty) and their likes are considered vulgar English words, and are therefore standardized as per rule 1 of Section III-F.

4) The rest of the Latin numerals are rarely used in SMS text and therefore no specific rules are designed to handle them.

J. Interchangeable Words

Some words in Romanized Urdu are interchangeable like 'aashiyan' and 'aashyaanah' [shelter], and therefore possess exactly same meaning. In such case, if both words are used frequently in Urdu, then both are considered as separate standardized words in the language vocabulary, otherwise the word that is rarely used is replaced by the one that is frequently used in the language by the human expert.

K. Differently Pronounced and Spelled Words

Some words in Urdu are pronounced differently when compared to the way that they are written. Such words when standardized by the human expert in Romanized Urdu are spelled the way that they are pronounced rather than the way that they are written in Urdu, e.g. 'ambiya' [prophets] instead of 'anbiya'.

L. Common Words in Urdu and English

Some words are common in Urdu and English languages in such a way that they are pronounced the same way, have exactly same meanings but possess different spellings. Such words are standardized according to the rules governing their respective languages as mentioned in Section III-B depending upon the way that they are written originally, as in 'jangal' and 'jungle'. This class of words usually includes nationalities of countries, like 'pakistani' and 'paakistaani'.

M. Typos

A typographical error or typo is a mistake made either due to slips of the finger as usually observed while thumb typing SMS messages on mobile phones, or due to spelling errors. All such typos are corrected using a standard English spell-checking list.

Words that do not go through any transformation according to the rule base mentioned in this section are left unstandardized. This constitutes construction of the first Romanized Urdu-English standardizer for SMS text, which is based on a semi-autonomous rule-based model that covers a wide range of domains of natural language processing and SMS text. We incorporate the multi-lingual SMS standardizer in the application of keyword extraction in the next section.

IV. SMS Standardizer for Keyword Extraction

Given a collection of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ and domain-based information (stopword list and standardization repository), a keyword extraction technique outputs the top K most descriptive terms from document d_i as its keywords. Due to the presence of 'noise' and substantial variability in SMS text, several preprocessing procedures are carried out for noise reduction and text standardization before any keyword extraction technique can be applied on the document collection. Initially, all documents in $\mathcal{D}$ are tokenized via space delimitation, yielding the initial set of vocabulary terms $\mathcal{T}_0$. Subsequently, the following preprocessing procedures are employed: (1) Punctuation removal (2) Symbol removal (3) Numeral removal (4) Lower-case transformation (5) Stemming using Krovetz stemmer [10] (6) Removal of stopwords using stopword list obtained from AutoMap software[8] (7) Standardization of text using SMS standardizer.

Notice that in general, each preprocessing procedure reduces the size of the resulting vocabulary (number of unique terms). This concept is discussed in detail in Section IV-A. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$ be the refined set of terms obtained after the final preprocessing procedure is implemented, where M is the size of the vocabulary. We now represent the documents in vector space of size M using the TFIDF technique as mentioned below:

$$TFIDF = \frac{count(t_j|d_i)}{\sum_j count(t_j|d_i)} \times \log \frac{N}{count(d_i|t_j)} \quad (1)$$

Using (1), each document $d_i \in \mathcal{D}$ is mapped into the vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]$, where $x_{ij} \geq 0$ is the weight of term t_j in document d_i. After this transformation, the entire document collection $\mathcal{D}$ can be represented by the $N \times M$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$. The top K keywords for document d_i are the top K terms in the document (in the ith row of the document-term matrix) with the highest term weights.

We evaluate keyword extraction performance through TFIDF on a smaller collection of 20,000 documents that are extracted randomly from the larger SMS collection utilizing which the standardization repository was built up (see Section I). We observe improvements in performance after each preprocessing procedure is applied on the new collection in order to determine the effectiveness of each procedure. The new collection, labeled as Collection-KE, is restricted to 20,000 documents due to the huge cost, time and effort incurred with manual labeling of documents for evaluation. We get this collection labeled by utilizing six student annotators, each of whom was assigned a unique sub-collection of documents. The annotators, each of whom is fluent in the languages present in the collection, were asked to extract a maximum of three keywords from each document due to the short length of SMS messages. The labels are eventually stemmed using Krovetz stemmer and standardized with the help of SMS standardizer in order to maintain consistency with the system generated keywords. The performance is evaluated by matching the output terms generated by TFIDF with those identified by human annotators using the standard F-measure (F1-measure) for the top 1, 3, and 5 keywords for each document in the collection.

A. Experimental Results

Figure 2 demonstrates keyword extraction performance in F-measure for $K = 1, 3$ and 5 obtained at different preprocessing procedures using TFIDF on Collection-KE. The key preprocessing procedures that improve keyword extraction performance significantly include punctuation removal, case transformation, stemming and standardization for all values of K. The use of the SMS standardizer produces the highest impact among all preprocessing procedures by improving keyword extraction performance by upto 23%, 14% and 15% in F-measure at $K = 1, 3$ and 5, respectively, thus establishing the significance of standardization as a preprocesing procedure for SMS text.

We also study the effect of different preprocessing procedures on vocabulary size of the collection. Initially the vocabulary size of Collection-KE is 50,741 terms, which after the final preprocessing procedure is reduced to 15,629 terms. Figure 3 illustrates the vocabulary size of

[8] http://www.casos.cs.cmu.edu/projects/automap/

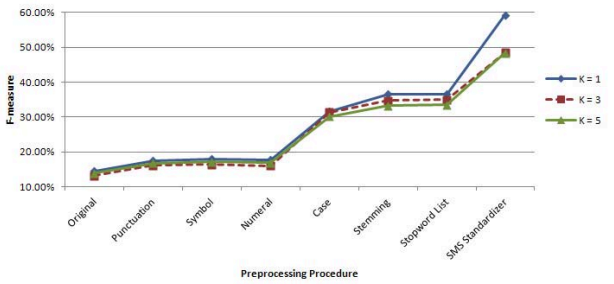

Fig. 2: Keyword extraction performance at different preprocessing procedures

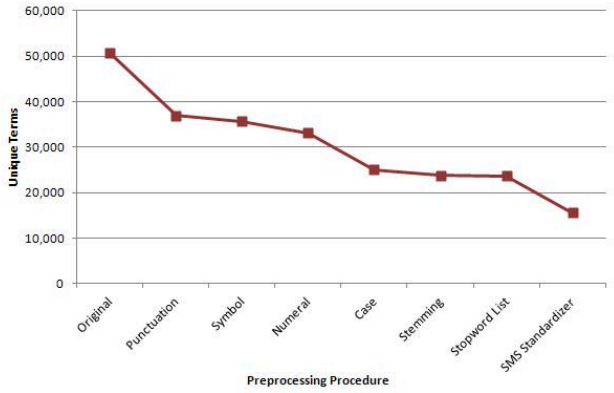

Fig. 3: Number of unique terms at different preprocessing procedures

Collection-KE after each preprocessing procedure is carried out. It is observed that major reductions in vocabulary size are achieved after punctuation removal, case transformation, and implementation of SMS standardizer, with the latter two reducing the vocabulary size by 16% each which are only next to that achieved by the punctuation removal procedure.

It is interesting to note that the observed improvements in keyword extraction performance correlate strongly with the reduction in unique terms at each preprocessing procedure. From this analysis we infer that preprocessing of SMS text results in significant noise reduction and text standardization in documents which cause reduction in unique terms accordingly that in turn results in substantial improvement in keyword extraction performance.

V. Conclusion and Future Work

In this paper we present a rule-based model for normalization of SMS text that is multi-lingual in nature and contains data from both morphologically and non-morphologically rich languages, i.e. Urdu and English respectively. The SMS standardizer is built upon a sophisticated set of rules that tackle challenges related to noise and multi-varied composition of SMS text, besides dealing with the complexities of multi-lingualism, MRLs, non-MRLs, and Latin script. We utilize the standardizer in the application of Keyword Extraction from SMS text, where it produces the highest impact (among several preprocessing procedures) on improvement in keyword extraction performance by upto 23% in F-measure.

The developed standardizer builds a repository of NSWs alongwith their unique standard forms from a large SMS collection. We plan to devise a technique to automatically map new NSWs obtained from documents from a newer or larger collection to the standardized forms already present in the repository. One way of carrying out this task would be to use a term weighting technique based on syntactic features due to the non-applicability of semantics in multi-varied microblogs. Another solution could be the development of a joint model of transliteration (from Romanized Urdu to Urdu, using the SMS standardizer) and machine translation (from Urdu to English, using a novel technique). We also intend to evaluate the SMS standardizer on other microblog collections like Twitter and Facebook in order to extend the scope and applicability of our work.

References

[1] "Nastaliq script," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Nastaleeq

[2] W. Wu, B. Zhang, and M. Ostendorf, "Automatic generation of personalized annotation tags for twitter users," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10, pp. 689–692.

[3] A. Aw, M. Zhang, J. Xiao, and J. Su, "A phrase-based statistical model for sms text normalization," in *Proceedings of the COLING/ACL on Main conference poster sessions*, ser. COLING-ACL '06, pp. 33–40.

[4] C. Kobus, F. Yvon, and G. Damnati, "Normalizing sms: are two metaphors better than one?" in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08, pp. 441–448.

[5] D. Pennell and Y. Liu, "Normalization of text messages for text-to-speech," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, march 2010, pp. 4842 –4845.

[6] D. Pennell and Y.Liu, "Toward text message normalization: Modeling abbreviation generation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, may 2011, pp. 5364 –5367.

[7] R. Beaufort, S. Roekhaut, L.-A. Cougnon, and C. Fairon, "A hybrid rule/model-based finite-state framework for normalizing sms messages," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10, pp. 770–779.

[8] F. Liu, F. Weng, B. Wang, and Y. Liu, "Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, ser. HLT '11, pp. 71–76.

[9] B. Han and T. Baldwin, "Lexical normalisation of short text messages: makn sens a #twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11, pp. 368–378.

[10] R. Krovetz, "Word sense disambiguation for large text databases," Ph.D. Thesis, University of Massachusetts, Amherst, USA, 1995.