Rule Based Automatic Generation of Query Terms for SMS Based Retrieval Systems

Aarti Kumar
Department of Computer Applications
Maulana Azad National Institute of Technology
Bhopal, India

Sujoy Das
Department of Mathematics & Computer Applications,
Maulana Azad National Institute of Technology
Bhopal, India

Abstract : Every big and small need is fulfilled through the small hand held device called mobile. But matching the text standard of a mobile having limited space and of users' varying moods or limited capacity and knowledge, makes it difficult to comprehend what the user is actually seeking when it comes to processing strings sent through SMS text. This paper presents an overview of how strings sent through SMS can be processed in the simplest possible way and how they can be then used for matching best possible word from a standardized collection. In this paper we have presented an entirely different approach to handle the noise in the queries and have tried to bring forth a Rule Based Approach for automatic generation of query terms which converts a syntactically incorrect query to a semantically fruitful one.

Keywords: Domain dictionary; Synonym dictionary; phonetic replacement; rule based

1. INTRODUCTION

The advent of mobiles with all facilities like net connection, mobile banking, Bluetooth and wi-fi has brought the whole world into the palm of our hands. Every big and small need is fulfilled through this small hand held device. Blogs, Twitter, socializing sites and SMS are the most common approach now-a-days for all kinds of communication. These have limited space to convey a message. But all these are being used not only to communicate but also to seek information from various sources to help decision making. These sources, which are generally some organizations, government agencies or educational institutions, receive such strings and try to answer the query sent through any of such information seekers. Mostly when it comes to SMS text, the user behavior is not consistent and how they will frame their SMS depends solely upon their, attitude, mood, age and nature. In general it has been observed that the younger generation has started using somewhat cryptic text for sending their messages and communicating. The words that they use are non-standard and are not found in any of the standard dictionaries. But at the same time the users within themselves are very comfortable with that and comprehend such kind of text very well. Problem arises when the same has to be understood and analyzed by an online system which relies only on a standard dictionary and refuses to recognize such words or texts. These systems expect the query in a proper standard language format. The challenge is to process such deformed strings, find out meanings from such words which otherwise grammatically appear meaningless, match them with the original query and supply the answer. This involves a lot of steps and lots of calculations.

The approaches that are generally adopted to handle such queries are either cognitive which exploits the human brains to answer questions or they simply use the principles of Natural language processing or else they are IR based which treats the question-answering as a problem of Information Retrieval and employs various rules to search the corpora of text and find the correct match to help answer the question in the best possible manner.

In this paper we have tried to bring forth a Rule Based Approach for automatic generation of query terms. This approach for finding the best question answer pair for a given query, from a list of FAQs, is easy to understand, handles the noisy SMSs efficiently and is simple to incorporate.

The rest of the paper is structured as follows: In Section 2 work done by different authors on SMS based FAQ retrieval are discussed. Section 3 discusses the proposed approach. Section 4 and 5 discuss the experiment performed and the observations respectively. Section 6 discusses the intricacies involved and the rules formulated and finally Section 7 presents conclusion.

2. RELATED WORK

Aw et al., 2006[14, Choudhury et al., 2007[15] and Kobus et al., 2008[13] have worked towards removing noise from SMS. An aligned SMS corpus and conventional language is required for training by the techniques employed by them. As reported by [4], Acharya et al., 2008 worked towards mapping non-standard words to their corresponding conformist recurrent form through an unsupervised technique. The algorithm proposed by Govind Kothari et al.[4] takes care of the noise in a SMS query by formulating query similarity over FAQ questions along with handling semantic variations in question formulation. As his approach considers it as a combinatorial search problem, therefore, the search space consists of combinations of all possible dictionary variations of tokens in the SMS query.

Deirdre Hogan et al.[5] presented a paper on SMS based FAQ retrieval in FIRE 2011. Their approach consists of first transforming the noisy SMS queries into a normalized, corrected form. The combined results of three different retrieval mechanisms are then used to retrieve a ranked list of FAQ results from the normalized

queries. The information gathered from retrieval results are then used for classifying and tagging out-of-domain (OOD) queries.

V. M. Pathak and M. R. Joshi[6] have worked on Marathi language and showed their results for Marathi language retrieval using SMS based query. They have used Vector space model and Cosine Similarity on ITRANS Marathi Literature documents to rank the documents as per their relevance for each selected query.

Paul Cook and Suzanne Stevenson in their work An Unsupervised Model for Text Message Normalization [7] have also given an unsupervised noisy-channel model for normalization of SMS text with 59% accuracy

Deana L. Pennell and Yang Liu [8] have given a system for normalizing text for Text-To-Speech engine. A classifier is used to form rules and to transform standard text to texting abbreviations. A reversal of the mappings gives the English words from these abbreviations. The intervention of human annotators is needed for the task of abbreviating.

Fei Liu et al.[9] have proposed a unified letter transformation approach where human supervision and pre-categorization of non standard words are not required. Nonstandard tokens under a sequence labeling framework have been generated from the dictionary words performed character-level alignment on a large set of noisy training pairs. They have reported absolute accuracy gain of 21.69% over deletion-based abbreviation system and of 18.16% over jazzy spell checker

3.PROPOSED APPROACH

Most the work that has been done towards processing such strings focuses on the SMS queries that are received and tries to solve the problem by processing these noisy queries using various methods and by adopting complicated processes. This involves quite a lot number of steps to solve a problem. But based on the fact that the number of SMSs received by any organization is, in majority of cases, manifolds the number of questions in the FAQ corpora, if instead of trying to mould the new generation language into standard language, it would be simpler if we do something which adapts to this language and moulds the standard language to harmonize with the current language in trend.

Although the works done by Deana L. Pennell and Yang Liu[8] and Fei Liu et al.[9] use the similar concept of processing the dictionary terms, but the work done in [8] requires human intervention, does not deal with substitutions of text which is very frequently seen in SMS text, does not deal with synonyms and is basically an interpretation of text only for text to speech conversion and for query formulation.

Again the proposed system by Fei Liu et al. [9] is very complicated involving many steps and complex calculations.

The proposed work is a step towards the same and uses very simple but efficient procedure to solve the problem. No human intervention is involved in normalizing the text and no complex procedures have been used. The idea is that instead of processing the umpteen numbers of SMSs received, process the terms of the standard query, which is limited in number, to match the terms of SMSs text.

Through this approach it will be less time consuming, less intricate and easier to resolve the queries and will also require lesser number of steps.

4. EXPERIMENT AND ANALYSIS

To come out with the rules and verify it an experiment was conducted on three batches of MCA students with 90 students per batch of our institute and on a group of middle aged people of officers club. They all were given to write on a piece of paper any message of their choice that they would like to send their family and friends keeping in mind the restriction of 150 characters imposed by mobile SMS service. They were asked to write it in the way they would write it in an SMS. Not restricting them to a domain made it easy for us to get SMS terms on a variety of topic and helped us in forming the rules. Further data was collected from comments of Face Book.

After forming the rule we applied the rules on dataset provided to us by FIRE 2012 for SMS based FAQ retrieval. The dataset consisted FAQs in three languages – English, Hindi and Malyalam that were collected from online sites - both government and private. They included domains like Railway Enquiry, Telecom, Health and Banking .The SMSs for FIRE were generated by asking college students to write down their information need using a mobile phone. We used the English dataset of 999 KB of SMSs for our study to generate the original FAQs and to retrieve the answer. 610 SMS terms of FIRE were also analyzed. Table 1 and Figure 1 show the analysis of rules incorporated in SMSs of FIRE 2012 English SMS dataset.

Windows was the platform used for testing algorithm. Java (jdk1.7.0_07) was used to implement the algorithm through program and test the results for original FAQs. Rule numbers 5.1 to 5.6 were incorporated in the programs to generate terms and to retrieve the queries to get the results. The testing is still in progress and is giving convincing results.

Table 1. Analysis of rules incorporated in SMSs of FIRE 2012 English SMS dataset

| Rule Applied	No. of instances	Percentage out of 610 words
Acronym	27	4.43
Repeating letter dropped	1	0.16
H following w dropped	5	0.82
Digit replacement	4	0.66
Vowels completely dropped	154	25.25
Partial dropping of vowel	171	28.03
First few letters to represent word	40	6.56
Combination of rules	51	8.36
Last g dropped	2	0.33
Stylistic variation	70	11.48
Phonetic replacement of word	10	1.64
Phonetic replacement of substring	19	3.12
Repeating letter dropped	7	1.15
Spelling mistake	39	6.39
Complete standard words	10	1.64
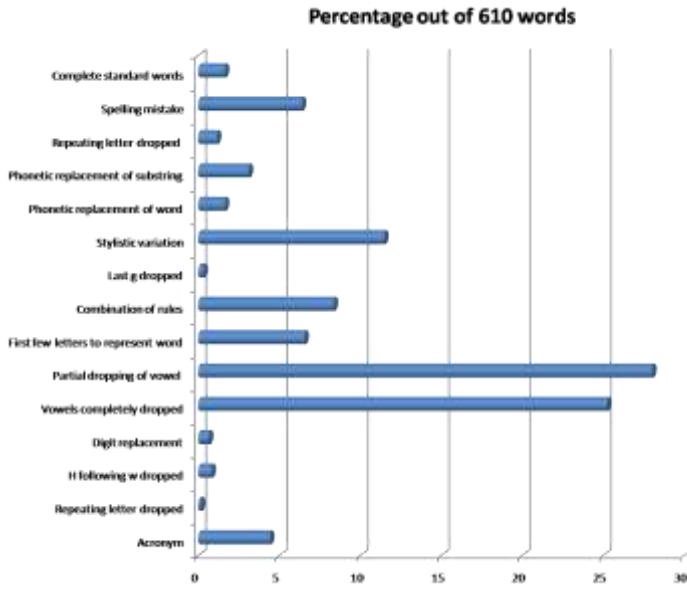
Figure 1: Graph showing percentagewise analysis of Rules incorporated in SMSs of FIRE 2012 English SMS dataset

5. OBSERVATIONS

After going through many SMSs it was observed that although the users are of varied backgrounds, behaviors, likings, moods and many a times also unsure of spellings and sentence formations, yet they are aware of one thing and that is phonetics and limited space. They try to convey their message with their full meanings to the ones they wish to communicate. For this they have to distort, shorten the words and use phonetic or digit substitutions more often. Still there are some general rules that are followed while creating SMSs. Eight such rules have been recognized which are as follows:

5.1 Omitting Vowels Partially or Retaining Vowels Partially (Generally the First or the Last Vowel)

If space allows, and the standard words contain more than one vowel, then only the first vowel from the words is removed to write the SMS term. Table 2 shows some common SMS terms conforming to this rule.

Table 2. SMS terms created by partially retaining or omitting vowels

SMS text	Standard word	SMS text	Standard word
ntion	nation	evaluation	valuation
bettr	better	xpect	expect
unending	nending	gmes	games
inactive	nactive	alwys	always
bke	bike	anarchy	narchy
latr	later	lbow	elbow

5.2 Omitting Vowels Completely

The SMS term is actually the consonant frame which is obtained by removing all the vowels from the standard term (Table 3).

Table 3. SMS terms created by fully omitting vowels

SMS text	Standard word	SMS text	Standard word
plc	place	wrst	wrist
strng	Stong, string	frm	from
fr	for	strtd	started
f	of	wht	what
bst	best	wrld	world
I	eye	cn	can
u	you	X	axe
whr	where	bt	But, bet
cn	can	nt	Not, net
fnd	find	rqts	raquets
hw	how	whch	which
s	is	rsdnt	resident
p	Pea	R	are
m	am	knw	Know, knew
srv	serve	gt	Get, got
srvc	service	B	Bee, be
wrn	warn	t	Tea, tee
thnk	thank		

5.3 Omitting Consecutive Repeated Alphabets

As per this rule, the letters appearing more than once in succession, are removed from the standard term to get the SMS term (Table 4)

Table 4. SMS terms created by omitting consecutive repeated alphabets

SMS text	Standard word	SMS text
Tenis	tennis	committee
comite	running	runing

5.4 Dropping Last 'g', Last 't', Last 'e' and also Dropping 'h' Associated with 'i' and 'w'

In most cases the last 'g' and't' are dropped if that does not affect the meaning of the word. The 'h' which follows 'w' or precedes 'i' is also dropped to form the SMS term. Table 5 shows some examples of SMS terms which are created in this manner.

Table 5. SMS terms created by 'g', last 't', last 'e' and also dropping 'h' associated with 'i' and 'w'

SMS	Standard	SMS	Standard

text	word	text	word
carin	caring	goin	going
jus	just	wat	what
usn	using	wch	which
comin	coming	wite	white
warin	wearing	hav	have

5.5 Representing Words by Digits, Single Alphabets or Combination of Alphabets

One or more alphabets, digits, special characters or symbols are used to create a full word with the same phonetics (Table 6).

Table 6. SMS terms created by representing words by digits, symbols, single alphabets or combination of alphabets

SMS text	Standard word	SMS text	Standard word
a	Answer,	n	An, and
b	Be, bee	o	Owe,oh, ough(tho)
c	See, sea,she	q	Queue, question
d	the	u	you
i	Eye	v	we
mt	empty	ne	any
y	why	5	five
1	one	6	Six
2	To, two, too	7	Seven
3	Three	8	Eight
4	For,four	9	nine
10	ten	+	plus
&	and	ur	your
@	at		

5.6 Using Alphabets, Symbols and Digits to Replace Substring Character, Substring Bi-gram, Substring Trigram and Substring Quad-gram Instead of the Whole Word

This rule is like the previous rule but instead of replacing the whole standard word, its substring is replaced. The single character, bigrams, trigrams and quad grams which are part of a string(standard word) and sound like single alphabets, symbols and digits, are replaced by the corresponding alphabet, symbol and digit to get the SMS term (Table 7).

Table 7. SMS terms created by replacing Substring character, substring Bi-gram, substring Trigram and substring quad-gram by alphabets, symbols and digits

Substring	Single Alphabet/ symbol/digit	Examples
si	c	cmple

di, th,de	d	dat, der, d@, dtel, d,dm,dem
gee, gi	g	
ck	k	quik
oh,ow	o	ro, o
et, eat, eight	8	gr8, p8, 8y,
ks	X(thanx)	
que(st)	q	qst
oo,o	u	u, gud, cum
wee	v	vd,vp
whi	y	yl
se, s, st,is	z	uz, juz, balz,whatz
ine,in,line	9	f9, f9d, on9(online)
an	n	ny
and	&	gr&, b&,br&
at	@	h@, w@
one, on	1	1s,1c, 1ce,up1
to,	2	2maro
fore	4	b4, der4
ha	a	av

5.7 Represent words using combination of the above rules

The same person at different time can frame the same word in a different form. This mood based generation of words generally applies the combination of two or more of the above mentioned rules. This variation may also depend upon the age of the users. Table 8-13show some examples of SMS created combining more than one rule.

Table 8. SMS words created using rule A and F

SMS text	Standard words	Explanation
Wrstb&	wristband	1. Partial vowel dropped 2. 'and' replaced by symbol '&'
B4	before	1. Vowels dropped 2. 'fore' represented by '4'

Table 9. SMS words created using rules B and D

SMS text	Standard words	Explanation
wch	which	1. Vowel dropped 2. 'h' associated with 'i' dropped
wt	what	1. Vowel dropped 2. 'h' following 'w' dropped

Table 10. SMS words created using rules B and F

SMS text	Standard words	Explanation
frndz	friends	1. Omitting vowels 2. Replacing 's' with 'z'
plz	please	1. Vowel removed 2. 's' replaced by 'z'

Table 11. SMS words created using rules B and C & also rules C and F

SMS text	Standard words	Explanation
clg	college	1. Dropping vowels 2. Dropping repeated 'i'
balz	balls	1. Repeating letter dropped 2. 's' replaced by 'z'

Table 12. SMS words created using rules D and F

SMS text	Standard words	Explanation
watz	whats	1. Dropping 'h ' associated with 'w' 2. Replacing 's' with 'z'
w@	what	1. 'h' associated with 'w' removed 2. 'at' replaced by symbol @

Table 13. SMS words created using 3 rules. First 3 columns show the example which uses rule A, C and D and the last 3 columns show the example which uses rules B, C and F

SMS text	Standard words	Explanation
caln	calling	1. Partial vowel removal 2. Removing repeated letter 3. Dropping last 'g'
10s	tennis	1. Vowel dropped 2. Repeating letter dropped 3. Phonetic replacement by digit

5.8 Abbreviations as SMS terms

Many standard and custom made abbreviations of the standard words are used to create the SMS (Table 14).

Table 14. Standard and custom made abbreviations as SMS terms

SMS text	Standard Forms	SMS text	Standard Forms
Comfy	comfortable	Btw,	between
		b/w	
lab	laboratory	audi	auditorium
motivatn	motivation	std	standard
no	number	envt	environment
Tomm, morrow	tomorrow	hol	holiday
mgmt	management	govt	government
nite	night	cum	come
id	identity	Bcoz, cuz	because
Diff/diffce	difference	pic	picture
Bday, b/day	birthday	ello	hello
exam	examination	chap	chapter
2maro	tomorrow		

5.9 Stylistic variation

This is the trend among the youngsters who use any style of writing a SMS term and these cannot be predicted. Table XV shows some examples of such stylistic variations.

Table 15. Stylistic variations used for SMS terms

SMS text	Standard word
Mah, ma	my
betta	better

6. GENERAL RULES FOLLOWED FOR CREATING AN SMS

It was also observed that code for the same word varied from person to person. Where one coded tomorrow as '2maro, the other coded it as '2morrow' and yet another as 'tomorrow'. Yet the variations followed one or the other rules from rule number 1-8. So for obvious reasons a rule based system, which is developed on the basis of the above rules, will certainly take care of all the morphological variations of the query terms.

It was observed that the dropping of partial vowels(the first one generally) when a vowel starts a word in many cases gives a totally mismatched reduced structure of word which does not match the standard word at all and does not convey the phonetics as well. For example, on one hand where 'awful', 'own', 'idle' etc will give totally meaningless word, on the other 'ideal' will give a totally different word with a different meaning. Also dropping the vowels completely sometimes poses problems and gives the same consonant skeleton for different words. For e.g. 'Quiet' and 'quite' and 'quit' in such a case will give the same term 'qt' which naturally will not be coded as same in the SMS terms. Likewise, 'form', 'from' and 'forum' also be coded as the same term 'frm' which is actually a SMS term for 'from'. For such terms the proximity phrase search or context searching needs to be done.

Each SMS S is a collection of tokens S = t1, t2, t3…., tn. Let Q denote the set of questions in any organization's FAQ corpus. Each question Q* ϵ Q a also a collection of tokens. The algorithm below

aims at finding the question Q* from corpus Q which is the best equivalent for SMS S.

The proposed approach takes into account the above mentioned rules to develop a rule based approach for query expansion and proceeds through the following steps for automatic generation of query Q*:

1. Developing a Domain dictionary consisting of all terms that appear in standard collection of queries in the FAQs of the organization.
2. Creating a synonym dictionary for all the terms of the domain dictionary and mapping them to the terms in the domain dictionary.
3. Processing the terms of the domain dictionary and their synonyms, using the above mentioned rules, for creating a rule based list of variants that could match the SMS text and arranging them along with the standard term in alphabetical order
4. Searching the SMS terms in the list of generated variants
5. If a match is found, mapping the matched term to the domain dictionary term
6. Retrieving the actual query term
7. Retrieving all the queries which contain that particular term
8. Repeating the process for all terms of SMS and generating a query list for each
9. Applying 'project' operation to find the query which is common in all the query lists
10. This will give the required query.
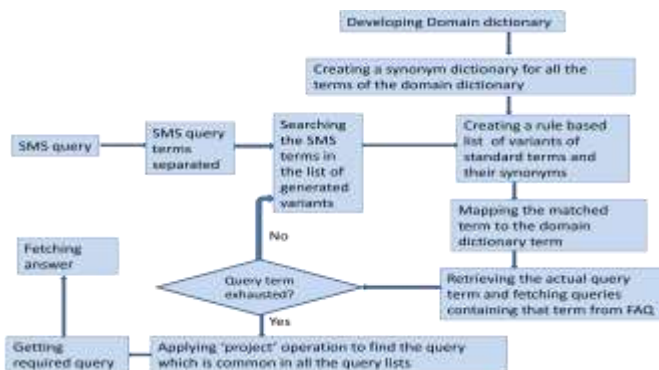
Figure 2 describes the steps for query expansion.

Figure 2. Steps for Rule Based Expansion of Query Terms

Also taking into account the problems with dropping of vowels and after carefully analyzing a long list of words, it was observed that the following conditions have to fulfilled before removing vowels either partially or completely to avoid generation of totally junk words. In all these cases if after removing the vowel if the reduced word sounds the same, then the rule can be relaxed.

1. If two lettered word, vowel should not be removed unless the removal does not affect the phonetics. . E.g. *to, no, of* should be retained as it is but in *in* it can be removed to give the same sound even with *n.*

2. If three lettered word with only one vowel e and that too as the last letter, vowel should not be removed. E.g. *the* should remain as it is or the whole word should be replaced by d
3. If three or five lettered with single vowel (a,i,o) as the middle alphabet
4. If four lettered word and last letter s, vowel should not be removed. E.g. tabs
5. If four lettered word and last letter is the second vowel, last vowel should not be dropped
6. If four lettered word and last two letters same, vowel should not be removed. E.g. call, tall (removal gives cll, tll which will also be obtained in case of the words cell and tell)
7. If I and e appear in a word with an intervening consonant, then i is replaced by y and e is dropped e.g. Bike->byk, like->lyk, time->tym, life->lyf
8. Same way if o and e appear in a word with an intervening constant, then o is replaced by u and e removed e.g. some->sum, come->cum
9. If the word starts with a vowel then first vowel should not be removed. E.g. if not followed, idle will give dle which is a junk word and our will become ur which is the abbreviated form of your
10. Two vowels appearing together in succession (except for a and u appearing with o or e appearing with u) vowel should not be removed.
11. If a and o or u and o appear together in succession and a follows o or u follows o then a and u can be removed but o has to be retained. E.g. fought(foght), goal(gol)
12. If u and e appear together and e follows u then u can be removed. Eg. Guess
13. If u comes after o with a consonant in between then u has to be retained, o can be removed. E.g. forum-> frum
14. If ou come together and are the only two vowels of the word followed by l, o and l can be removed e.g. should ->shud, could->cud
15. S occurring in between the words should not be replaced by z
16. If word with single vowel and last letter is g, g should not be removed.
17. If word is a standard abbreviation it should be preserved as it is.
18. The abbreviated months and week days should be retained as it is.
19. Last two letters same and single vowels before that, vowels should not be removed. Eg still, stall, wall, mess
20. If vowels alternate with consonant and there are only two vowels in the word, then the initial vowel can be removed. Eg forum-> frum, daring->dring
21. If applying two rules ie of vowel removal and of removal of one of the two consecutive repeated character, if the word after the first rule gives the ending as nng then the rule for removal of last g should not be applied as the third rule to generate abbreviations e.g. training-> trainng-> traing-> train×

7. CONCLUSION

It has been observed that dropping a vowel is the most obvious and used rule for creating an SMS term. The reason being its presence in every word more than once and dropping it considerably saves space where space is a constraint. As the proposed system takes into account the inadvertent but unanimously accepted vocabulary of the cryptic terms used in the SMS and tries to form rules from the same, it is able to devise a very easy and less time consuming method for automatic generation of expanded query terms for SMS based retrieval system. It does not involve any complex calculation and uses only table look ups and mappings to expand the query terms. Also it uses lesser number of steps to do the same. The unordered, unplanned and non standard format of the SMS will take the standard form automatically when this Rule Based system will be applied and will certainly prove a step forward in this direction. However, this approach does take into account the typographical errors and spelling mistakes. The experimentation is still in its progress stage. Once this proves successful, it can be deployed on a larger corpus of FAQs.

8. ACKNOWLEDGEMENT

9. REFERENCES

[1] Manning Christopher D., Raghavan P. and Schulz H. An Introduction to Information Retrieval, Cambridge University Press.

[2] Baeza-Yates R. and Ribeiro-Neto B. Modern Information Retrieval, Pearson education.

[3] Joel S. and Samuel W. Information Retrieval System Design for Very High Effectiveness.

[4] Kothari G. et al. SMS based interface for FAQ retrieval.

[5] Hogan D. et al. SMS Based FAQ Retrieval.

[6] Pathak V. M. and Joshi M. R. Itransed Marathi Literature Retrieval Using SMS Based Natural Language Query.

[7] Cook P. and Stevenson S. An Unsupervised Model for Text Message Normalization

[8] Pennell Deana L. and Liu Y. Normalization Of Text Messages for Text-To-Speech.

[9] Liu F. et al. Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision.

[10] Frakes W.B. and Baeza-Yates R. editors. 1992. Information Retrieval: Data Structures and Algorithms. Prentice-Hall.

[11] Salton G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Massachusetts.

[12] Ravino B. and Boll R. 1993. A Natural Semantics for Information Retrieval. In East Pacific Rim Symposium on Applied Linguistics, pages 161–163.

[13] Kobus C., Yvon F. and Damnati G. 2008. Normalizing SMS: are two metaphors better than one? In Proc. of the 22nd Int. Conf. on Computational Linguistics, pp. 441–448. Manchester.

[14] AiTi Aw et al. 2006. A phrase based Statistical model for SMS text normalization. In Proc. of the COLING/ACL 2006 Main Conference Poster sessions, pages 33-40, Sydney.

[15] Choudhary M. et al. 2007. Investigations and modeling of the structure of texting language. International Journal of Document analysis and Recognition, 10(3/4):157-174.

[16] Rebecca E. et al. 2001. Y do tngrs luv 2 txt msg. In Proceedings of 7th European Conference on Computer Supported Cooperative Work, pages 219-238, Bonn, Germany.

[17] Sproat R. et al. 2001. Normalization of Non-standard words. Computer Speech and language, 15: 287-333.