SMS based FAQ Retrieval: A Theme Matching Scheme

Deba Prasad Mandal

&

Saptaditya Maiti

Machine Intelligence Unit

INDIAN STATISTICAL INSTITUTE

KOLKATA

email: *dpmandal@isical.ac.in*

ROADMAP

- Introduction
- Motivation
- String Similarity Measures
- Proposed Theme Matching Scheme
    - Preprocessing (FAQ & SMS Queries)
    - Query Matching
    - Relevance Decision
- Implementation & Result
- Conclusions

Short Messaging Service (SMS)

- A low cost, easy and immediate mode of communication
- High reach capability
- Used for
  - Personal messages
  - Enquiry
  - Commercial purpose
- Being increasingly used as a source of information
- Texts are noisy

Noise in SMS

- Mainly due to
  - Keypad constraints on mobile devices
  - Maintain the limitation of characters (160 characters)
  - Poor language Skill

A. Non-intentional

- Commonly used Abbreviations [*e.g.: Math*, *Max*, *SBI, don't*]
- Spelling errors
- grammar mistakes

B. Intentional

- Non-standard Spellings [*e.g.: Trng* (*Training*), *Ppl* (*People*)]
- SMS specific Abbreviations [*e.g.: Prog* (*Program*), *Mob*(*Mobile*)]
- Phonetic Transliteration [*e.g.: 4get* (*Forget*), *Lyk* (*Like*)]
- Use of Latin Characters for native languages [*e.g.: Darun* (Excellent)]

Noise in SMS (Cont…)

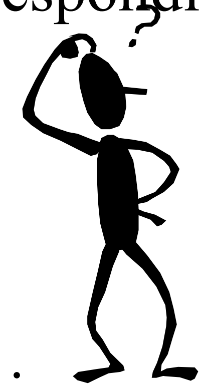
- Language used in SMS may be non-noisy for human communicators
- However, the words/characters used in such communication differ from standard language, and so they would be considered noise when processed by an automatic system/ tool

Frequently Asked Questions (FAQ)

- A useful source of information about an organization
- Contains listed questions and answers
- Compilations of information which are the result of certain questions constantly being asked
- Tries to keep answers to all the possible questions coming from users
- Sentences are noise free

SMS based FAQ Retrieval

- **What?**
  - Retrieving information from FAQ corpora corresponding to an SMS sent by user



- **Why?**
  - Growth of mobile telecommunication
  - Portability of a mobile device ensures information access from anywhere
  - Immediate and low cost services
  - High retention levels

Motivation

Some Typical FAQ Queries

- **What is the coverage offered by the Mediclaim Policy?**

  (*Mediclaim Policy*; *coverage*; *offered*)

- **If people had smallpox previously and survived, are they immune from the disease?**

  (*smallpox*; *immune*; *disease*; *survived*; *previously*)

- **Where can I find information about bulk repackaging of pesticides?**

  (*repackaging of pesticides*; *information*; *find*; *bulk*)

- **Why is it harder to get insurance if drivers in my household have bad driving records?**

  (*insurance; drivers; driving records*; *get*; *harder*; *bad*)

Motivation (Cont…)

Theme of a Query

- Nouns are found to have highest ability in reflecting/ representing the theme of a sentence/ query.
- This ability decreases for verbs, adjective-adverbs and other parts of speech.

Theme Matching Scheme

- Tries to find the Theme of FAQ queries (Noun terms
- The matching of the FAQ theme with an SMS query is checked. If checking is satisfactory, the matching of the full query is then checked

String Similarity Measures

Four similarity measures are applied for the matching of strings (with varying matching score).

Complete/Full Match

Both the strings are the same

Partial Match

A substring (*cash*, *cashless*)

Soundex Match

Similar sounding words (*person*, *prsn*)

Approximate Match

Limited letter mismatch (*passport*, *pport*)

Soundex Match

Soundex Algorithm [*O'dell, Russel*]

a) Retain first letter of the word and remaining letters are replaced by their codes

b) For the consecutive occurrence of the same digit, drop all but the first

c) Drop all '0's

d) Convert to the form 'letter digit digit digit' by dropping right most digits (if there are more than three digits) or by adding trailing zeroes (if there are less than three digits)

Letter	Code
A,E,I,O,U,Y,H, W	0
B,P,F,V	1
C,G,J,K,Q,S,X, Z	2
D,T	3
L	4
M,N	5
R	6

Instead of restricting to code size to **4** , we have taken the full code i.e., the step d) is modified as

d') Convert to the form 'letter digit digit …… '

KNUTH, D. E. Sorting and searching,Addison-Wesley, Reading, Mass.,1973.

Approximate Match

- For a given pair of strings, the best matched string is determined
- A similarity matrix $D^{m \times n} = [d_{ij}]$ is obtained as where

$$d_{ij} = 1 \quad \text{if } w1[i]=w2[j]$$

$$= 0 \quad \text{otherwise}$$

- A traversal algorithm along the '1' entries of *D* in the diagonal/right/down word directions is proposed starting from the (1,1) position

  Each traverse provides a matched string

  The string longest matched string (and have better lower order matched) is finally selected as the best matched string

Approximate Match: An example

- $w_1 =$ *photograph*; $w_2 =$ *photogap*

$$D = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

- Matched strings: ‘*p*’, ‘*ph*’, ‘*pho*’, ‘*phog,*’ ‘*phop*’, ‘*phoap*’, ‘*phogp*’, ‘*photog*’, ‘*phogap*’, ‘*photogp*’, ‘***photogap***’
- Best matched string = ‘***photogap***’

Approximate Match Score

Higher Positional weight (P_i) is considered for lower order letter matches (*e.g.*, P_i is 5,4,3,2 for i=1,2,3,4 respectively and P_i=1 for i>4)

- Matching Score, *S*, is then calculated as
-

$$S = \frac{\sum_{i=1}^{m} P_i . K_{1_i} + \sum_{i=1}^{n} P_i . K_{2_i}}{\sum_{i=1}^{m} P_i + \sum_{i=1}^{n} P_i}$$

where $Kj_i = 1$ *if the ith letter of the jth (=1,2) string is matched with the best matched string*

0 Otherwise

- *E.g.*, S (*photograph*, *photogap*)= *0.93889*

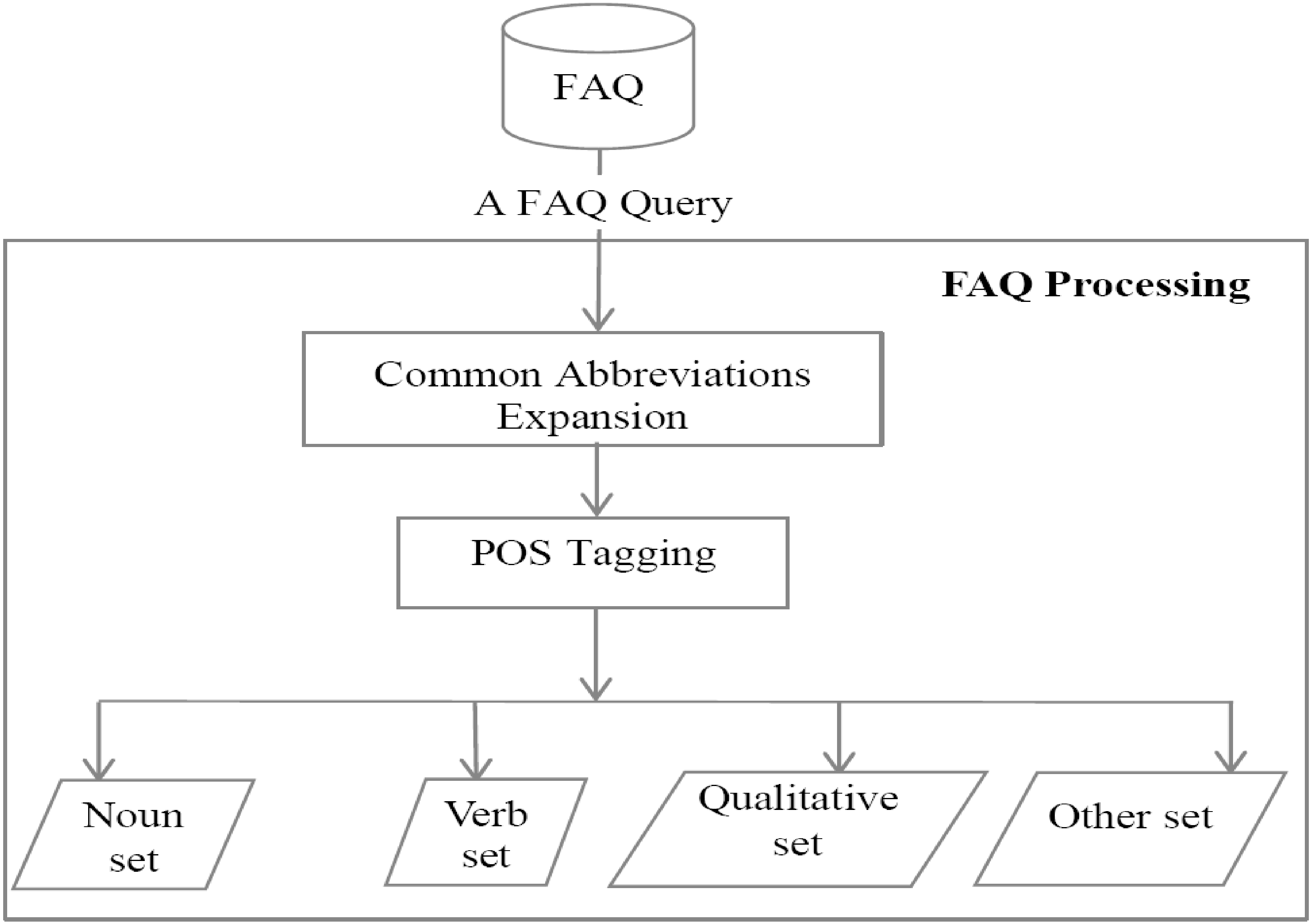
Compound Term

- A group of consecutive terms together carry a specific meaning which is usually different from each individual term
- Compound Nouns:
- Consecutive nouns (*e.g., Career counseling*)
- a noun preceeded by an adjective (*e.g., Prime Mininter*)
- a noun preceeded by an gerund verb (*e.g., Running water*)
- a preposition in between two nouns (*e.g., Master of Science*)
- Compound Adverbs:
- a Wh-adverb followed by an adjective (*e.g., How long*)
- Compound Term Match: If each individual term matches

Present Approach

- FAQ Processing
- SMS Query Processing
- Query Matching
- Relevance Decision

FAQ Processing



Common Abbreviation Expansions

- Linguistically valid abbreviations, if any, of the FAQ queries are replaced by their expanded forms
- Some Typical Examples:
  - *Subjects:* Math(s), Engg, Chem, Bio, ...
  - *Degrees:* BSc, BA, MCom, BTech, BBA, BCA, BEd, PhD, HS, ...
  - *Positions:* PM, IPS, CAO, ...
  - *Organizations:* Govt, SBI, RBI, Co, ...
  - *Cordial numbers:* 1st, 2nd, ...
  - *Verb conjugation and contraction:* I'm, you're, don't, haven't, won't, shan't, ...
  - *Others:* PC, TV, Exams, Ans, Qns, Acc, Max, Min, info, univ, ...

POS Tagging

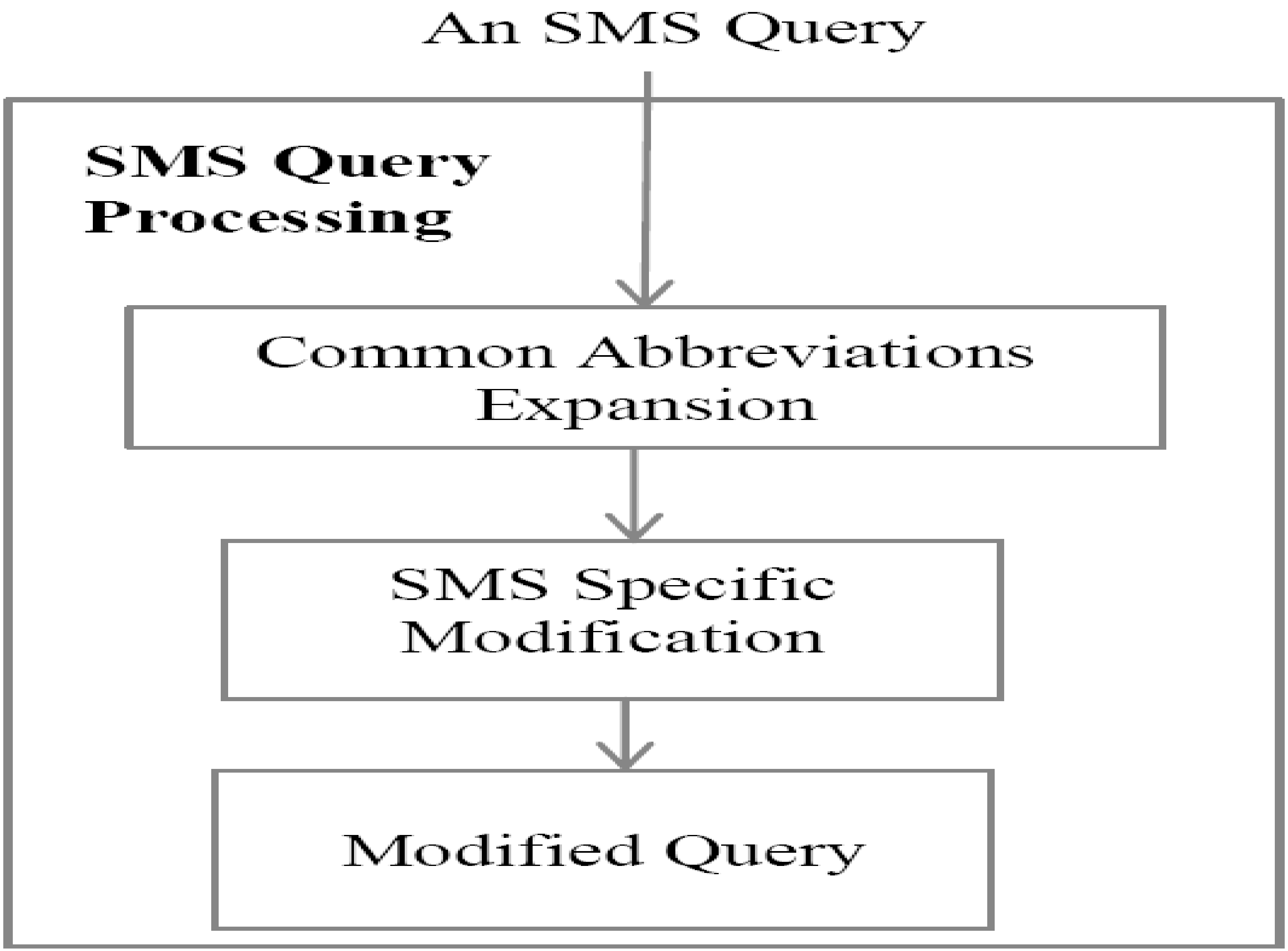
- Used Stanford POS Tagger
- It puts a POS Tag for each of the words in the FAQ queries
- Tags:

Noun:	NN, NNP, NNPS, NNS
Verb:	VB, VBD, VBG, VBN, VBP, VBZ
Qualitative:	JJ, JJR, JJS, RB, RBR, RBS
Others:	CC, CD, DT, EX, FW, IN, LS, MD, PDT, POS, PRP, PRP$, RB, RBR, RBS, RP, SYM, TO, UH, WDT, WP, WP$, WRB

- Compound Nouns & Compound Adverbs are identified
- Each FAQ query is decomposed into 4 term sets

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.* In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

SMS Query Processing



SMS Specific Modification

- Linguistically invalid abbreviations, are replaced by their expanded forms
- Some Typical Examples:

what:	*wht*, *wat*, *wt*, *vt*
what is:	*whats*, *wtz*, *vats*
which:	*wich*, *whch*, *wch*, *vich*, *wh*, *whc*
program:	*prog*
building:	*bldg*
available:	*avbl*
required:	*reqd*, *reqrd*
problem(s):	*prob*(*s*)
want to:	*wanna*
give me:	*gimme*
important:	*imp*
mobile:	*mob*, *mbl*

A Modified SMS Query

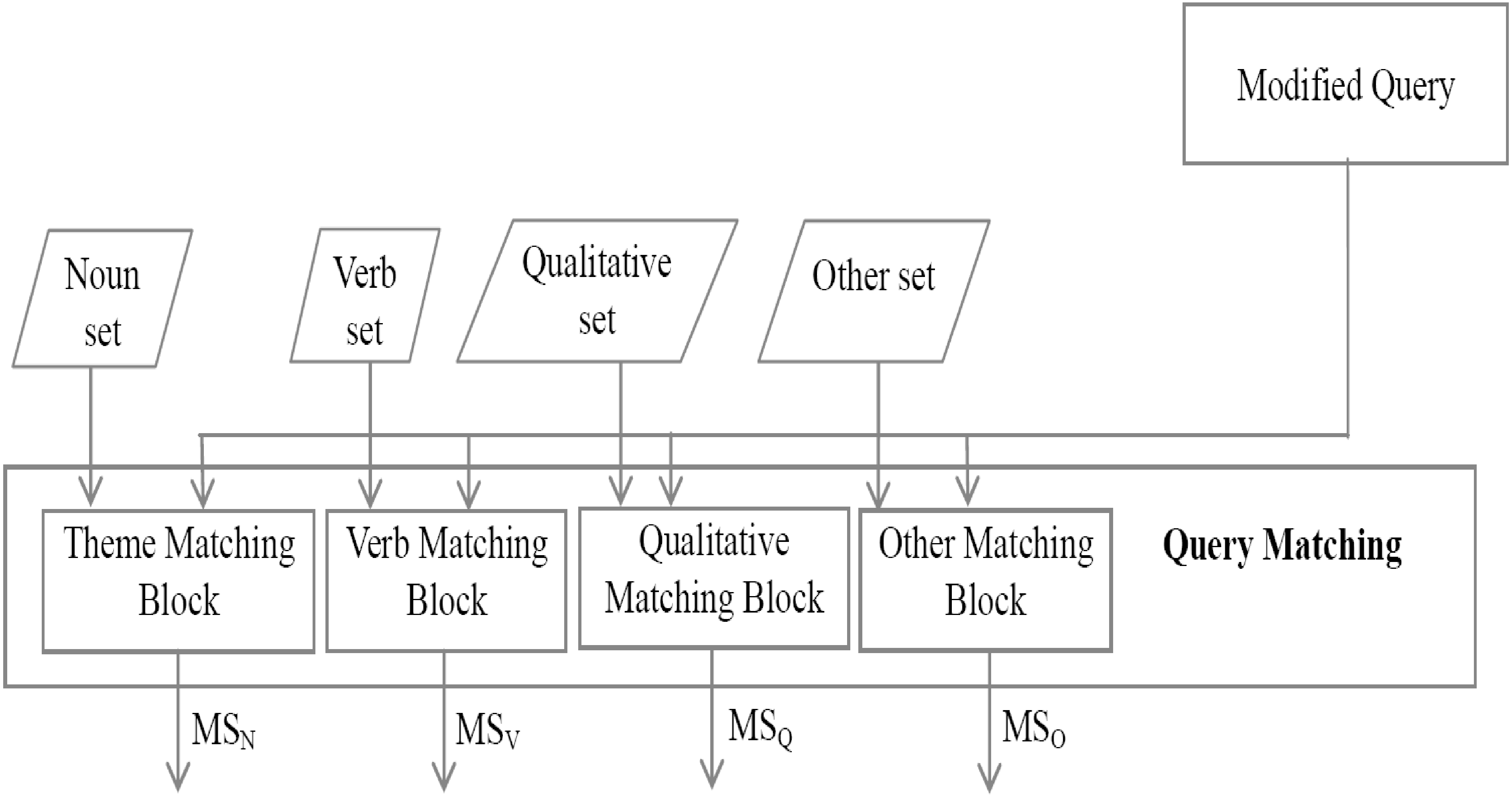
Query Matching

- Concerned with the quantification of the matching between the modified SMS query and each of the FAQ queries (4 term sets)

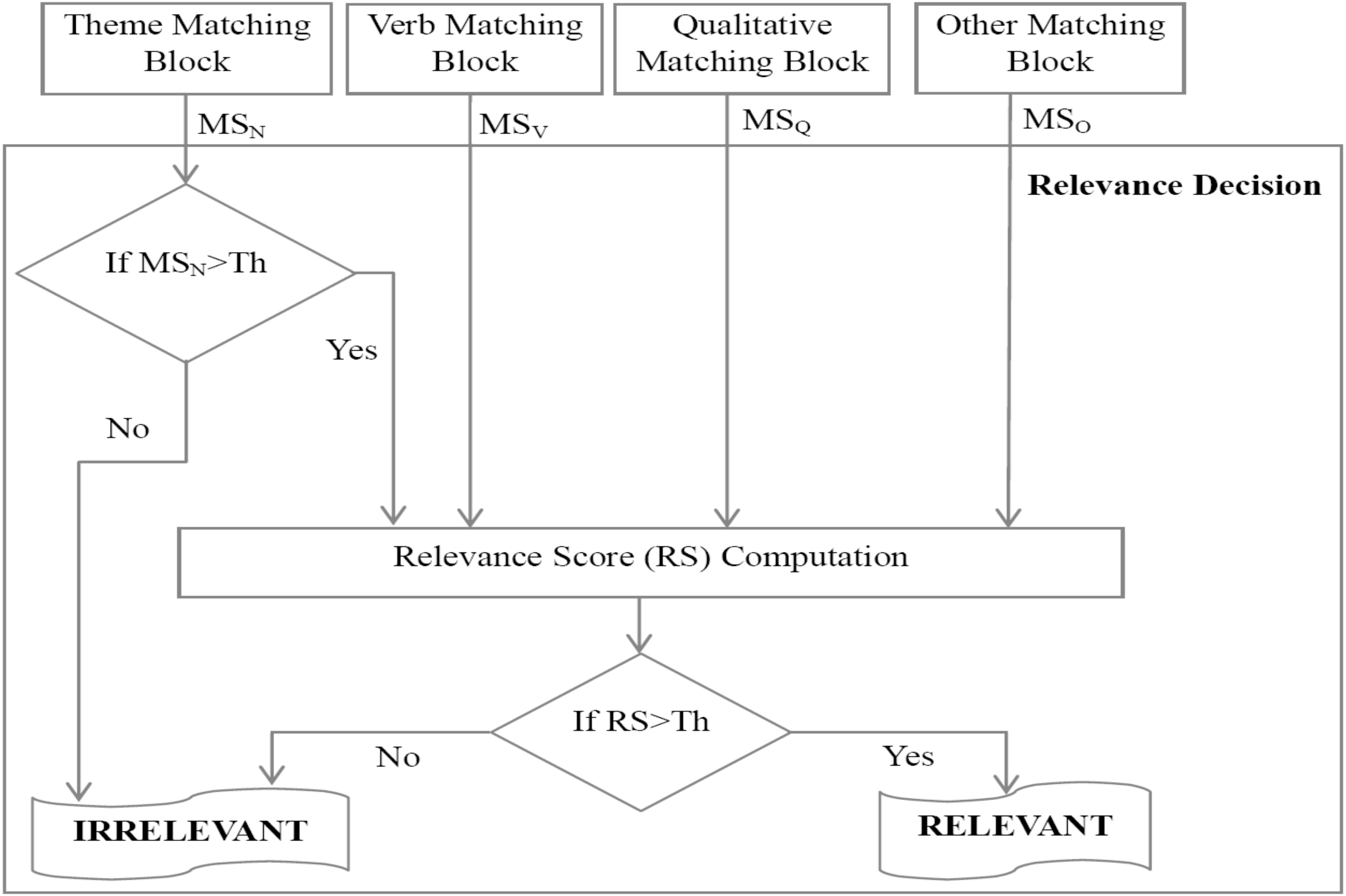
- Applied 4 Similarity Measures (Complete, Partial, Soundex & Approximate matches) sequentially

- Each similarity measure assigns a specific match value as
  - Complete Match :1
  - Partial Match : V_{pm}
  - Soundex Match : V_{sm}
  - Approximate Match: V_{ap} (defined earlier)

Query Matching (Cont….)



Relevance Decision



Relevance Decision (Cont….)

The four matching blocks of the Query Matching section provide the matching scores MS_N, MS_V , MS_Q and MS_O

Theme Verification: If $Average(MS_N) < Th$, the theme match is unsatisfactory and the FAQ query is rejected

- Otherwise Theme Match is satisfactory
  - Four significance factors $I_N > I_V > I_Q > I_O$ are considered
  - Relevance Score (*RS*) between the FAQ query (*q*) and SMS query (*s*) is determined as

$$RS(q,s) = \frac{I_N . MS_N + I_V . MS_V + I_Q . MS_Q + I_O . MS_O}{|s| - MS_O} \times T$$

Relevance Decision (Cont….)

- : $1/(|s| - MS_o)$ acts as the Length Normalization Factor

  [As (|*s*| - *MSo*) is the maximum possible match between *s* & *q*]

- *T* acts as the Size Mismatch Penalty which is defined as

$$T = \begin{cases} \frac{|q|}{|s|} & if\ |q| < |s| \\ \frac{|s|}{|q|} & if\ |q| > |s| \\ 1 & if\ |q| = |s| \end{cases}$$

- If ***RS(s,q) > Th***, *q* is considered to be relevant to *s*

Otherwise *q* is irrelevant to *s*

Relevance Decision (Cont….)

- Output:
  - Relevant Set: All relevant FAQ queries in order of relevance scores are decided as the relevant set
  - NULL: In case all the FAQ queries are irrelevant

Implementation

- FIRE 2012 SMS-based Monolingual English FAQ Retrieval Task
- Dataset
  - 7251 FAQ queries from different domains including Railways Enquiry, Telecom, Health, Banking, GK, Career counseling etc.
  - 1733 SMS queries (726 'In Domain' and 1007 'Out of Domain')
- Constants of the Proposed System
  - Threshold value: $Th = 0.3$
  - Matching constants: $V_{pm} = 0.5$, $V_{sm} = 0.8$
  - Significance factors: $I_N = 1$, $I_V = 0.8$, $I_Q = 0.5$, $I_O = 0$

Implementation: An Example

SMS query	can i take a policy for mre dan 1 year							
	FAQ Query	**Tagged FAQ**	Nouns	Verb	Qualitative	Others	SMS Length	Query Length
	Can I take a policy for more than one year	can_MD i_FW take_VB a_DT policy_NN for_IN more_JJR than_IN one_CD year_NN	Policy; year	take	more	can; I; a; for; than; one		
Total score	3.2		2	1	0.8	4	10	10
Normalized score	0.533		policy 1.0	take 1.0	more 0.8	can		
Penalty score	1		year 1.0			a		
Final score	**0.533**					for		
						i		
	Can a policyholder with 1 year no claims bonus have open driving on their policy	can_MD a_DT policyholder_NN with_IN 1_CD year_NN no_DT claims_NNS bonus_NN have_VBP open_JJ driving_VBG on_IN their_PRP$ policy_NN	policyholder; 1 year; claims bonus; policy	driving	open	can; a; with; no; have; on; their;		
Total score	3		3	0	0	2	10	15
Normalized score	0.375		1 1.0			can		
Penalty score	0.667		year 1.0			a		
Final score	**0.25**		policy 1.0					

Results

Queries	In Domain	Out of Domain	Total
No of queries	726	1007	1733
Correct	686 (*0.9449*)	988 (*0.9811*)	1674 (*0.965955*)
MRR	–	–	0.963754

Conclusions

- **Proposed a theme matching scheme for SMS FAQ Retrieval**
  - The FAQ queries are decomposed into four term sets (noun, verb, qualitative, others) with the help of a POS Tagger
  - Nouns are considered to represent the theme of a query
  - An FAQ query is considered to be relevant to an SMS query if the theme matching score as well as the relevance score are both satisgfactory
  - The output for an SMS query is NULL (‘Out of Domain’) if all the FAQ queries are found to be irrelevant
- **A new approximate string similarity measure is proposed**
- **Performance of the proposed system is very much dependent on the accuracy of the POS Tagger**

Questions?

Thank You !!!