SMS BASED FAQ RETRIEVAL

Topics

- Introduction
- Improvements for English -
  - Merge consecutive Terms.
  - Stemming.
  - Score Calculation.
- Improvements for Hindi -
  - Normalization of FAQ and SMS.
  - Similar Characters Substitution
  - Stemming.

Introduction

Step-1: Pre-processing on SMS query & FAQ corpus.

- Stop words removal
- Creation of Domain dictionary & Synonym Dictionary

Step-2: For each *token in SMS*

- Find ranked list of FAQ terms present in Domain dictionary.

Step-3: Find the Candidate Set of questions.

Step-4: For each FAQ in *Candidate set*.

- Calculate Score.

Step-5: Return FAQs having highest score as a result.

Algorithms for SMS based FAQ Retrieval System

FIRE 2011

- Contributions in FIRE 2011.
  - Modified Score Calculation.
    - Score = Similarity Score + Length Score + Proximity Score.
  - Consider Answers in FAQ retrieval.
    - If there are many FAQs with similar score, then break the tie using answers.
    - Or If no matching FAQ found, then match with the answers to get result.

FIRE 2012

- Improvements for English language-
  - Merge consecutive Terms.
  - Stemming.
  - Improved Score Calculation.
- Improvements for Hindi language-
  - Normalization of FAQ and SMS.
  - Phonetic Character Substitution.
  - Stemming.

Merge Consecutive Terms

1. During preprocessing of the SMS text.
   - Merge SMS words.

2. While matching a SMS with FAQ.
   - Merge SMS words.
   - Merge FAQ words.

During preprocessing of SMS

- SMS :
  - Wat r symptms of smll pox ?
- After Joining SMS tokens :
  - Wat r symptms of smllpox ?

- Criteria to Join SMS terms t1 and t2 –
  - They must be consecutive.
  - Term similar to t1 or t2 does not exist in the Domain dictionary.
  - Term t3 exists in Domain dictionary, where t3=t1+t2.

Domain Dictionary

Term
Symptoms
Small
Smallpox
Home
Loan
Homeloan

For each SMS term Si {

If [! termExists(S_i) || ! termExists(S_{i+1})] && [termExists(concat(S_i , S_{i+1}))]

Si = concat(S_i , S_{i+1});

}

Joining consecutive SMS terms

- SMS-1: Hw to get *home loan* frm bank ?
- FAQ-1: How to obtain *homeloan* from bank ?

- During processing term 'home' & 'loan' are not combined because both terms exists in domain dictionary.

- Criteria to Join SMS terms S1 and S2:
  - Similarity(S1,F1) < similarity (S1+S2, F1).
  - eg. Similarity(home,homeloan)< similarity(homeloan,homeloan).

- Calculation of Weight:
  - $\omega_i = \alpha(S_i + S_{i+1} , F_i) * idf(F_i)$;
    - $\omega_i = \alpha($ home + loan , homeloan$) * idf($homeloan$)$;

Domain Dictionary

Term
Symptoms
Small
Smallpox
Home
Loan
Homeloan

Joining consecutive FAQ terms

- SMS-2: Hw to get *homeloan* frm bank ?
- FAQ-2: How to obtain *home loan* from bank ?

Domain Dictionary

Term
Symptoms
Small
Smallpox
Home
Loan
Homeloan

- Criteria to Join FAQ terms F1 and F2:
  - Similarity(S1,F1) < similarity (S1, F1+F2).
  - eg. Similarity(homeloan,home)< similarity(homeloan,homeloan).
- Calculation of Weight:
  - $\omega_i = \alpha(S_i, F_i + F_{i+1}) * \mathbf{Avg}[\ idf(F_i)\ ,\ idf(F_{i+1})\ ]$;
    - $\omega = \alpha(\ homeloan\ ,\ home + loan)\ * Avg[\ idf(home)\ ,\ idf(loan)\ ]$;

// Joining words while performing similarity calculation.

For each SMS token S_i {

For each FAQ token F_i {

// merge consecutive SMS words

$\alpha_1 = \alpha(S_i + S_{i+1} , F_i);\ idf_1 = idf(F_i);$

// merge consecutive FAQ words

$\alpha_2 = \alpha(S_i, F_i + F_{i+1});\ idf_2 = [\ idf(F_i) + idf(F_{i+1})\ ]\ /\ 2;$

// without merging

$\alpha_3 = \alpha(S_i , F_i);\ idf_3 = idf(F_i);$

$\omega_i =$ max($\alpha_1, \alpha_2, \alpha_3$) * <Corresponing IDF value>

}

}

Stemming

- Used Porter stemmer.
- Maximum similarity out of a Stem FAQ token and Non-stem FAQ token w.r.t. SMS token is considered.

- max[similarity(Stem FAQ token, SMS token) , similarity(FAQ token, SMS token)].

Score Calculation

Score $= W_1$*Similarity_Score$+ W_2$*Length_Score$+ W_3$*Proximity Score.

$$\text{Similarity\_Score}(Q) = \sum_{i=1}^{n} \max_{t \in Q \text{ and } t \sim s_i} \omega(t, s_i)$$

$$\omega(t, s_i) = \max(\, \alpha\,(t, si) * \text{idf}(t), \alpha\,(t\_stem, si) * \text{idf}(t)\,)$$

Length Score

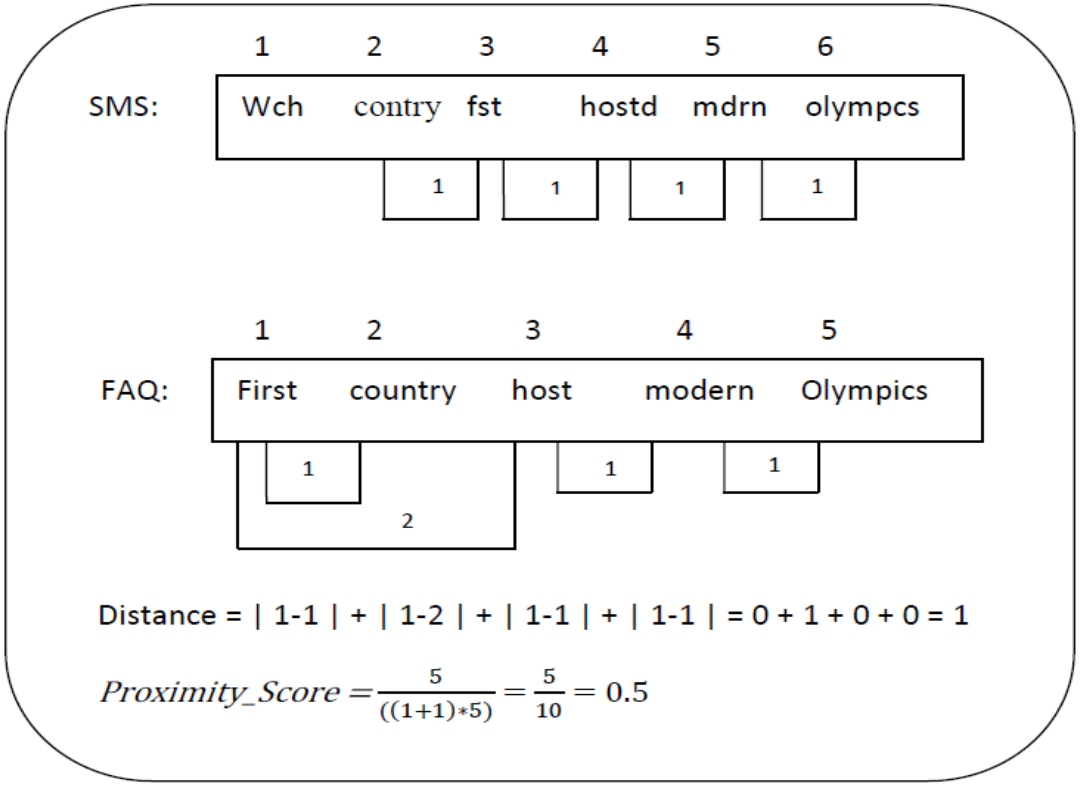
$$\text{Length Score} = \frac{2 * \text{MatchedTokens}}{\text{totalSmsTokens} + \text{totalFaqTokens} - 2 * \text{skippedFAQtokens}}$$

- *skippedFAQtokens* are the tokens skipped during the calculation of similarity score.
- These words are different from the Stop words.
- Skipped Words:
  - what, when, where, which, while, who, whom, why, will, with, would, yet, you, your…
- Stop Words:
  - a, an, and, are, as, at, be, but, by, for, of, on, or…

Proximity Score

$$\text{Proximity\_Score} = \frac{\text{matchedToken}}{((\text{distance} + 1) * \text{totalFaqTokens})}$$

$$\text{distance} = \sum_{k=0}^{n} \text{absolute difference between adjacent token pairs in SMS and corresponding pair in FAQ}$$



English Monolingual Task - Results

***** FIRE 2012 SMS TASK EVALUATION REPORT *****

No. of In-domain Queries :	726
No. of Out of Domain Queries:	1007
In Domain correct:	686/726 (0.94490355)
Out of Domain correct:	986/1007 (0.979146)
Total Score:	0.9648009
Mean Reciprocal Rank (MRR):	0.9711889

Hindi Monolingual Task

- **Normalization of FAQ and SMS**
  - Replace some characters to make the FAQ corpus and SMS queries consistent.
  - In most of the cases these characters does not change the meaning.
  - Normalization is done as a part of FAQ preprocessing and SMS preprocessing.

Normalization

- replace(़,)
- replace(ै,े)
- replace(ौ,ो)
- replace(ि,ी)
- replace(ु,ू)
- replace(ृ,)
- replace(ई,इ)
- replace(ष,श)
- replace(न,ण)
- replace(उ,ऊ)
- replace(ॉ,ा)

Makes फ़ायदा & फायदा equal

Phonetic Character Substitutions

Substitutions	
प	फ
त	थ
श	स
क्ष	स
ष	स
ज	झ
ब	भ
क	ख
द	ध
ड	ढ
ग	घ
न	ण
उ	ऊ
र	ऋ
ङ	ञ
च	छ

Hindi Word Stemmer

- Hindi light stemmer is used to stem hindi terms.
- The stemmer removes plural, gender and case suffixes from nouns and adjectives
- Eg. आवश्यकताएँ after stemming becomes आवश्यकता.

Hindi Monolingual Task - Results

***** FIRE 2012 SMS TASK EVALUATION REPORT *****

No. of In-domain Queries :	200
No. of Out of Domain Queries:	379
In Domain correct:	186/200 (0.93)
Out of Domain correct:	376/379 (0.99208444)
Total Score:	0.97063905
Mean Reciprocal Rank (MRR):	0.9728009

Malayalam Language

- Updated stop words list
- No dictionary used.
- Score Calculation: Only similarity score is used.

Malayalam Monolingual Task - Results

***** FIRE 2012 SMS TASK EVALUATION REPORT *****

No. of In-domain Queries :	69
No. of Out of Domain Queries:	11
In Domain correct:	44/69 (0.6376812)
Out of Domain correct:	10/11 (0.90909094)
Total Score:	0.675
Mean Reciprocal Rank (MRR):	0.7613843

Conclusion

- Following techniques have improved the accuracy of SMS based FAQ retrieval system.
  - Stemming.
  - Merging consecutive FAQ and SMS words.
  - Normalization & Phonetic character substitution for Hindi language.
  - Revised Length score.
- These techniques can be used independently to take their advantage in FAQ retrieval.

References

- “SMS based Interface for FAQ Retrieval “ by Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkaesan T. Chakaravarthy, L. Venkata Subramaniam. 2009.
- English Stemmer - http://tartarus.org/martin/PorterStemmer/
- Hindi Stemmer - www.unine.ch/info/clef/
- WH words - http://www.englishclub.com/vocabulary/wh-question-words.htm
- Forum for Information Retrieval Evaluation(FIRE) - http://www.isical.ac.in/~fire/