Working Note – FIRE 2013

FAQ retrieval using noisy queries

Divyesh Sanjay Kothari
ISM DHANBAD

Abhinav Saraswat
ISM DHANBAD

Sarang Kapoor
ISM DHANBAD

Anjaney Pandey
ISM DHANBAD

Sukomal Pal
ISM DHANBAD

mailto:divyesh2506@gmail.com

ABSTRACT

This working note presents the results of participation in the SMS-based FAQ Retrieval task using noisy queries task at FIRE. For FIRE 2013, we submitted runs for monolingual English subtask. In our approach we normalised the SMS query in English and submitted to the search engine which returns the ranked list of FAQ queries. Our best experiment achieved a MRR of 0.90 for English Monolingual subtask.

INTRODUCTION

The note presents the subtask of English monolingual subtask of SMS-based FAQ Retrieval using noisy queries at Forum of Information Retrieval and Evaluation 2013. The task consists of correcting the incoming of SMS query (English) and retrieving the query and its answer from the FAQ database. The incoming query in English are written in noisy SMS text and contains many misspellings, abbreviations and grammatical errors. These queries are normalised and matched with the FAQ query database.

The working procedure is comprised of three distinct phases, pre-processing, normalisation of query and retrieval of ranked list.

PRE-PROCESSINGS

The first phase includes the functions required for the normalisation of query and retrieval from the FAQ database.

- Construction of dictionary of words from the FAQ database for normalisation of string tokens of a query.
- Formation of Index 'I' of the questions in the FAQ database using Indri tool.
- Construction of 3grams (of alphabets) as per the Algorithm B* used in the hybrid.

ALGORITHM B*

- For each word in the dictionary which we already have ,we find all the 3 consecutive occurrence (*3gram*) of the letters in the word removing the vowels while retaining the first the letter as it is.

- We then map each of these 3grams to the original dictionary.

- These *3grams* are stored as 26 x 21 x 21 files ranging from abb to zzz (i.e. excluding the vowels for 2nd and 3rd letter), wherein each file consists of all the words of the dictionary in which the particular *3gram* occurred.

NORMALISATION OF QUERY

ALGORITHM : HYBRID

```
Require: SMS QUERY Corpora Q
Open Query Corpora Q
    for Every SMS query q ϵ Q tokenize q into words
        for Every word ϵ words
            if ( size ( word ) < 3 )
                do
                {
                    foreach ( dictionary word )
                    do
                    if ( min > EditDistance ( query word, dictionary word ) )
                    {
                    Min ← EditDistance ( query word, dictionary word ) ;
                    CorrectWord ← dictionary word;
                    }
                }
            else
                do
                {
                  Apply algorithm A* to get correct word ( CorrectWord )
                }
            FinalQuery ← FinalQuery + CorrectWord ;
        end for
            Retrieve the best 5 results using the previously built index for
                FinalQuery
    end for
```

Algorithm A*

e.g.(reservation)

1. Remove the vowels from the query word, i.e., *rsrvtn*
2. Take all the words from the files of trigram database *rsr, srv, rvt* & *vtn.*
3. Now for each word in *rsr* calculate the count of that word in all the files.
4. The word with highest count becomes the correct answer and if there are 2 words with the same count ,we consider the first word, which is output as the *CorrectWord*.

For the normalisation of query, hybrid approach is used which works on the string tokens of the query. It takes string tokens one by one and process it according to the size of the token. This procedure returns the query in the form of collection of string tokens with proper dictionary words.

RETRIEVAL OF RANKED LIST

For indexing and retrieval of query, indri tool (lemur project) library is used. The normalised query is submitted to this tool which returns the FAQID of top 5 matches from the FAQ database.

RESULT FOR ENGLISH MONOLINGUAL TASK

The text only result for English mono task :

```
***** FIRE 2013 SMS TASK EVALUATION REPORT *****

No. of In-domain Queries :200
No. of Out of Domain Queries:99

In Domain correct:174/200 (0.87)
Out of Domain correct:0/99 (0.0)

Total Score: 0.5819398

Mean Reciprocal Rank (MRR): 0.90007937

ERRORS:
```

The overall result for English mono task :

```
***** FIRE 2013 SMS TASK EVALUATION REPORT *****

No. of In-domain Queries :392
No. of Out of Domain Queries:148

In Domain correct:177/392 (0.4515306)
Out of Domain correct:0/148 (0.0)

Total Score: 0.32777777

Mean Reciprocal Rank (MRR): 0.47031212

ERRORS:
ENGLISH_NR_2013a_1438
ENGLISH_NRC_2013a_4118
ENGLISH_NRC_2013a_4123
```

Refrences

DCU@FIRE2012: Monolingual and Crosslingual
SMS-based FAQ Retrieval
Johannes Leveling
Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin 9, Ireland
johannes.leveling@computing.dcu.ie

SMS based Interface for FAQ Retrieval
Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarth, L. Venkata Subramaniam
IBM India Research Lab
lvsubram@in.ibm.com

Foundations of Statistical Natural Language Processing
Christopher D. Manning and Hinrich Sch¨ utze
(Stanford University and Xerox Palo Alto Research Center)

Angell, R.C., Freund, G.E. & Willett, P. (1983)
Automatic Spelling Correction using a trigram similarity measure

A spelling correction program based on a noisy channel model
MD Kernighan, KW Church, WA Gale