Normalization of Chinese Informal Medical Terms Based on Multifield Indexing

2014.12.8

Xie Zhongda

Outline

* Introduction
* Motivation
* Approach
* Evaluation
* Conclusion & Future Work

Introduction

* Healthcare data mining and business intelligence are attracting huge industry interest in recent years.

* People suffer from informal medical terms when applying data mining tools to textual medical records.

Motivation

* Many medical terms in the healthcare records are different from the standard form, which are referred to as informal medical terms in this work.

Informal Term	Standard	English Explanation
上感	上呼吸道感染	upper respiratory tract infection
TNB	糖尿病	diabetes
GXB	冠状动脉硬化性心脏病	coronary arteriosclerotic cardiopathy
Guillian-Barre氏综合征	古兰-巴雷综合征	Guillian-Barre syndrome
急性烂尾炎	急性阑尾炎	acute appendicitis

Motivation

* Study indicates that in Chinese healthcare records, a majority of the informal terms are abbreviations or typos.

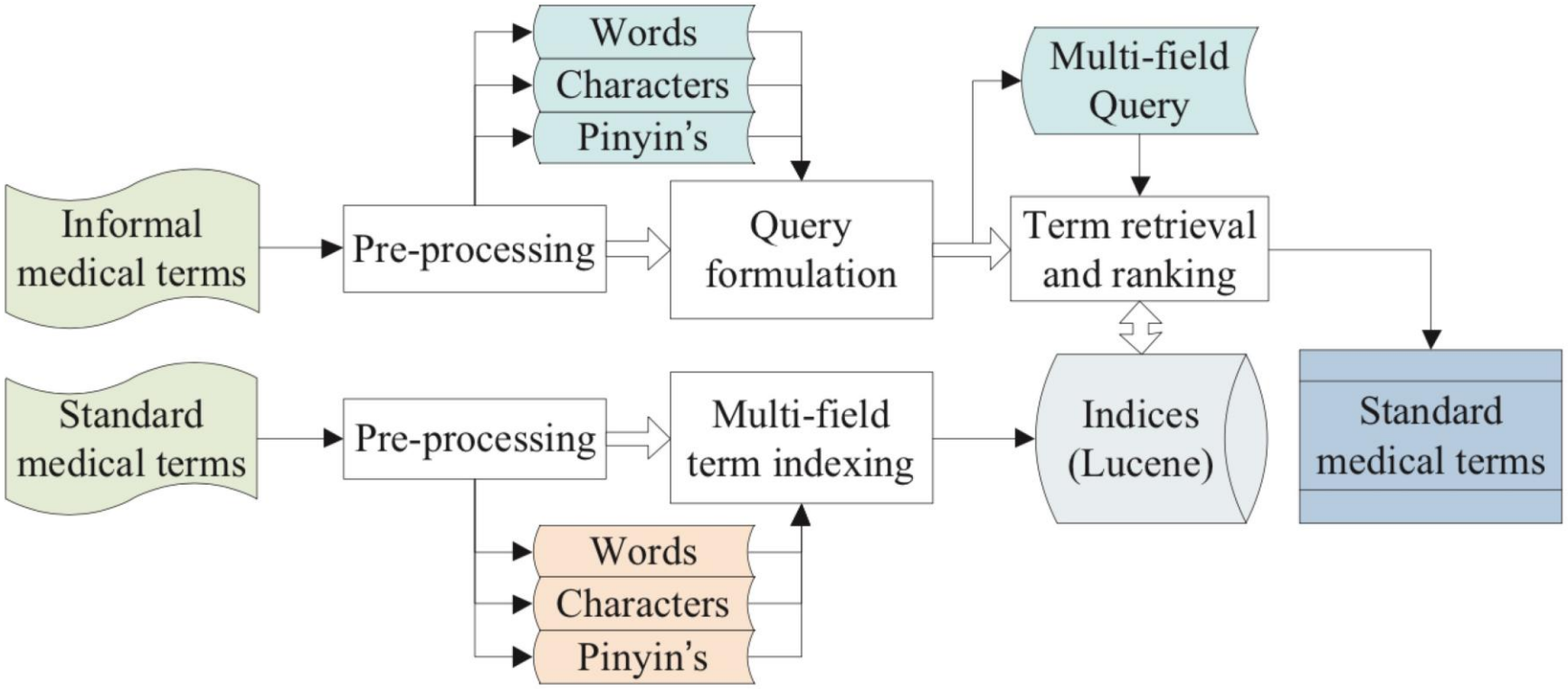
* Thus, targeting at the two types of informal medical terms, we propose a multifield indexing approach, which is able to normalize the informal medical terms.

Approach(1/3)

* We adopt the information retrieval framework and accomplish the medical term normalization task via term retrieval and ranking

* Information retrieval framework with four level indices: word, character, pinyin and initial.

Approach(2/3)


Words
Characters
Pinyin's
Multi-field Query
Informal medical terms
Pre-processing
Query formulation
Term retrieval and ranking
Standard medical terms
Pre-processing
Multi-field term indexing
Indices (Lucene)
Standard medical terms
Words
Characters
Pinyin's

Approach(3/3)

* Multi-field Term Indexing
  * Different weights
* Term Retrieval and Ranking
  * BM25 rank model

Index	Standard	Inform	Weight
Words	上呼吸道 感染	上 感	1
WordInitials	上 感	上 感	0.1
Pinyins	shang hu xi dao gan ran	shang gan	3
PinyinInitials	s h x d g r	s g	0.1
PinyinFinals	ang u i ao an an	ang an	2
Characters	上 呼 吸 道 感 染	上 感	1

Evaluation(1/3)

* Setup
  * Dataset
    * 300 pairs of informal medical terms and their standard counterparts
    * 125 Chinese abbreviations
    * 48 pinyin abbreviations
    * 127 typos
  * Evaluation Metric
    * P@N
    * Execution time

Evaluation(2/3)

* Experiments
  * Normalization Methods
    * Edit distance (EDDis)
    * Multi-filed cosine similarity (MSim)
    * Our Method(IRNorm)
  * The Fields in the Index
    * Using different fields in the index

MethodID	Word	Character	Pinyin
IRNorm-A	Y	N	N
IRNorm-B	Y	Y	N
IRNorm-C	Y	N	Y
IRNorm-D	Y	N	Y

Evaluation(3/3)

* Results

Methods	P@5	P@10	Time(milliseconds per term)
EDDis	0.748	0.762	120
MSim	0.853	0.892	180
IRNorm	0.892	0.907	6*

MethodID	P@5	P@10
IRNorm-A	0.398	0.412
IRNorm-B	0.624	0.653
IRNorm-C	0.685	0.723
IRNorm-D	0.892	0.907

Conclusion & Future Work

* Contributions of this work
  * Multiple fields make term normalization more accurate.
  * The normalization process is made much faster under the information retrieval framework.
* Future work
  * Explore how context helps to normalize the informal medical terms.
  * Develop the informal term detection algorithm, which can find inform terms automatically.

Thank you!