Normalization of Chinese Informal Medical Terms Based on Multi-field Indexing

Yunqing Xia[1], Huan Zhao[1], Kaiyu Liu[3], and Hualing Zhu[2]

[1] Department of Computer Science,
TNList, Tsinghua University, Beijing 100084, China
yqxia@mail.tsinghua.edu.cn
[2] Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong
hzhaoaf@ust.hk
[3] Information Networking Institute
Carnegie Mellon University, Pittsburgh, PA 15213, USA
liukaiyu1991@gmail.com
[4] Asia Gateway Healthcare Information Technology (Beijing) Co Ltd
Beijing 100027, China
hualing.zhu@aghit.com

Abstract. Healthcare data mining and business intelligence are attracting huge industry interest in recent years. Engineers encounter a bottleneck when applying data mining tools to textual healthcare records. Many medical terms in the healthcare records are different from the s-tandard form, which are referred to as informal medical terms in this work. Study indicates that in Chinese healthcare records, a majority of the informal terms are abbreviations or typos. In this work, a multi-field indexing approach is proposed, which accomplishes the term normalization task with information retrieval algorithm with four level indices: word, character, pinyin and its initial. Experimental results show that the proposed approach is advantageous over the state-of-the-art approaches.

Keyword: Term normalization, medical terms, indexing, ranking.

1 Introduction

Healthcare data mining and business intelligence are attracting huge industry interest in recent years. Healthcare industry is benefited from data mining applications for various parties [1] by feeding the demand of efficient analytical methodology for detecting unknown and valuable information in health data. With data mining tools, fraud in health insurance can be detected, medical advices can be provided to the patients at lower cost. Moreover, it becomes possible with data mining tools to detect causes of diseases and identify novel medical treatment methods. On the other hand, healthcare researchers may make full use of business intelligence tool in making efficient healthcare policies, constructing drug recommendation systems, or developing health profiles of individuals.

Engineers encounter a bottleneck when applying data mining tools on textual healthcare records. Many medical terms in the healthcare records are different

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 311–320, 2014.

from the standard form, which are referred to as informal medical terms in this work. We first give some examples in Table 1.

Table 1. Examples of informal medical termsp

#	Informal term	Standard Counterpart	English explanation
*E*1	上感	上呼吸道感染	upper respiratory tract infection
*E*2	TNB	糖尿病	diabetes
*E*3	GXB	冠状动脉硬化性心脏病	coronary arteriosclerotic cardiopathy
*E*4	Guillian-Barre氏综合征	吉兰-巴雷综合征	Guillian-Barre syndrome
*E*5	急性烂尾炎	急性阑尾炎	acute appendicitis

The examples in Table 1 actually presents the following three categories of informal medical terms:

- Abbreviation: The abbreviations can be further classified into Chinese word abbreviation, pinyin abbreviation and mixed abbreviation. For example, *E*1 ('上感', *shang*4 *gan*3) is a Chinese word abbreviation, *E*2 (TNB) is pinyin abbreviation and *E*3 (GXB) is a pinyin abbreviation for '冠心病(*guan*4 *xin*1 *bing*4)', which is a Chinese word abbreviation for '冠状动脉硬化性心脏病(*guan*4 *zhuang*4 *dong*4 *mai*4 *ying*4 *hua*4 *xing*4 *xin*1 *zang*4 *bing*4)'.
- Transliteration: The standard term is a transliteration of a word. In example *E*4*p*, '吉兰-巴雷(*ji*2 *lan*2 - *ba*1 *lei*2)' is transliteration of *Guillian-Barre*.
- Character input error: Some characters are wrong but phonetically equal / similar to the standard character. In example *E*5, '阑(*lan*2)' is replaced by 烂(*lan*4). This is typically caused by Chinese pinyin input tool.

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 multi-field indexing approach, which is able to normalize the informal medical terms with under the information retrieval framework with four level indices: word, character, pinyin and initial.

The standard medical terms are first segmented using the standard Chinese lexicon, namely, no medical terms is included. For example, term '上呼吸道感染(*upper respiratory tract infection, shang*4 *hu*1 *xi*1 *dao*4 *gan*3 *ran*3)' is split into {上呼吸道(*upper respiratory tract*)|感染(*infection*)}. As many abbreviations are generated based on word, we detect boundary of the words for the purpose to discover the word level abbreviations, e.g., *E*1 in Table 1. With the fine-grained words, we are also able to handle input errors as well as synonyms, e.g., ' 症(*symptom, zheng*4)' and '病(textitdisease, zheng4)'.

We also handle the medical terms on character level. For example, we split '糖尿病(*diabetes, tang*2 *niao*4 *bing*4)' into {糖|尿|病}. This makes it possible that we recognize TNB as its abbreviation. Again, the character level treatment is also useful to handle input errors, e.g., *E*5 in Table 1.

We further handle the medical terms on pinyin level. Every Chinese character holds a pinyin, which indicates how the character is produced. The purpose is to recognize the abbreviations that are comprised of initials, e.g., $E2$ in Table 1, or English word, e.g. $E5$ in Table 1.

We adopt an information retrieval framework in medical term normalization. We first index the words, characters, pinyin's and their initials with Lucene[1]. Using the input informal terms as a query, we then apply standard $BM25$ algorithm[2] to retrieve and rank the standard terms in multiple fields. Experimental results show that the proposed approach is advantageous over the state-of-the-art approaches.

The reminder of this paper is organized as follows. The related work is summarized in Section 2. The proposed method is described in Section 3. We present the evaluation as well as discussion in Section 4. We finally conclude this paper in Section 5.

2 Related Work

This work is related to two categories of previous work: medical term normalization and language normalization.

2.1 Medical/Biological Term Normalization

Medical term normalization is a major task in a few bio-medial natural language processing competitions, e.g., 2013 ShARe/CLEF eHealth Shared Task (Disorder Normalization in Clinical Notes) [2] and NTCIR11 Medical NLP Shared Task [3] (an ongoing task). The former task focuses on English terms while the latter on Japanese. There is no report so far on other languages. Unified Medical Language System (UMLS) is a commonly used English medical knowledge base [3, 4] which contains over 2 million names for some 900,000 concepts from more than 60 families of biomedical vocabularies, and includes 12 million relations among these concepts. Most recent work on English medical terms are based on UMLS [5, 2].

In the ShARe/CLEF eHealth Shared Task, acronyms/abbreviations are expected to be mapped to standard terms in UMLS. Researchers have developed systems to normalize acronyms/abbreviations in clinical texts for information extraction, information retrieval, and document summarization applications. Wu et al. (2012) compare performance of some some current medical term normalization tools, e.g., MetaMap, MedLEE, and cTAKES. It is showed that f-scores of these tools range from 0.03 to 0.73 [6].

As showed in the aforementioned work, lexicon based matching is a dominant solution. In this work, we base our work on the Chinese medical terminology system developed by Asia Gateway Co. Ltd. and explore advanced technology for term matching.

[1] http://lucene.apache.org/

[2] http://en.wikipedia.org/wiki/Okapi_BM25

[3] http://mednlp.jp/ntcir11/#task

2.2 Language Normalization

Language normalization is an important task for many natural language processing systems such as machine translation, information extraction and information retrieval. The problem is defined as detecting the so-called informal words from text and mapping them to their counterparts in the standard lexicon. In [7], Sproat et al. (2001) propose a ngram language model for this purpose. In [8], Xia et al. (2006) propose a phonetic model for chat language normalization.

The common characteristics of the above research lies in that they handle free natural language text. This work is different because we handle medical terms which cannot be directly mapped to any term in the standard medical lexicon. Our input is not the free text, but the terms that are already detected as non-standard ones.

3 Method

3.1 The Workflow

We adopt the information retrieval framework and accomplish the medical term normalization task via term retrieval and ranking. The general workflow is given in Fig 1.

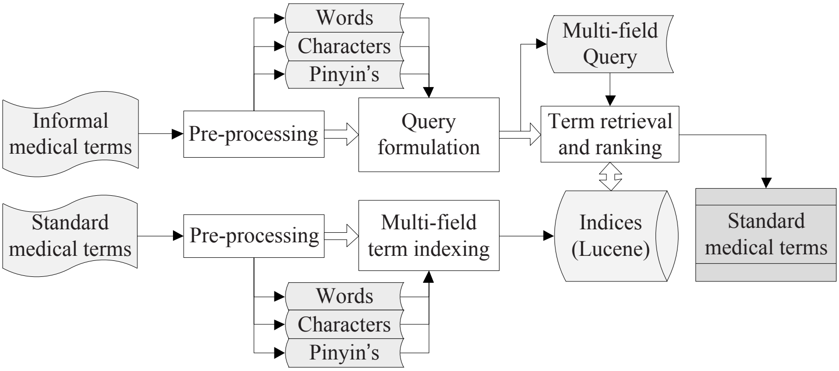

Fig. 1. The general workflow for medical term normalization.

The input to the system is medical terms which cannot be matched in the current ontology. The output is the standard counterpart terms. The horizontal part of the workflow is designed for standard term indexing, in which standard terms are first converted into sequences of words, characters and pinyin's, respectively. The words, characters and pinyin's are in turn input to Lucene to build a multi-filed index. The vertical part of the workflow is designed for informal term normalization. Based on standard $BM25$ score, we are able to retrieve and rank the relevant standard terms. The workflow is elaborated as follows.

3.2 Multi-field Term Indexing

Observation shows that informal medical terms are created with variations in word, character and pinyin levels, no matter deliberately or not. Enlightened by this, we propose to create index for the standard medical terms with multiple fields so that the given informal term can be matched within the standard terms according to relevance in word, character and pinyin levels.

Applying ICTCLAS[4], we first segment the input informal term in to words. Note that only standard vocabulary is used in ICTCLAS, thus a medical term is usually segmented into a few word. For example, '冠状动脉硬化性心脏病(*guan*4 *zhuang*4 *dong*4 *mai*4 *ying*4 *hua*4 *xing*4 *xin*1 *zang*4 *bing*4)' is segmented into {冠状动脉(*coronary artery*)| 硬化(*sclerosis*)| 性(*type*)|心脏病(*heart disease*)}.

According to unicode scheme, i.e. UTF-8, for the Chinese characters, we further split words into characters. For example, the above medical term is segmented into {冠|状|动|脉| 硬|化|性|心|脏|病}.

We further apply pinyin annotator, i.e. pinyin4j[5], to transcribe every Chinese word with pinyin. For example, pinyin annotation of '冠状动脉' is *guan4 zhuang4 dong4 mai4*. Note that the tool works better when a whole word is input because there are some Chinese characters which can be mapped top different pinyin's in various contexts.

There are some cases that English words/characters are contained in the Chinese medical terms. We use natural delimiter to split the English words/characters. For example, 'Guillian-Barre氏综合征{*Guillian-Barre syndrome*, *Guillian-Barre shi*4 *zong*1 *he*2 *zheng*4}' in example $E4$ in Table 1 is split to {Guillian|-|Barre| 氏| 综| 合| 征}.

The obtained words, Chinese characters and pinyin's are used to create a multi-filed index with the following structure:

```
INDEX = {
    string Words;\\Words in the term to be indexed
    string WordInitials;\\The first characters of the words
    string Pinyins;\\Pinyin's of the term
    string PinyinInitials;\\Initials of the above pinyin's
    string PinyinFinals;\\Finals of the above pinyin's
    string Characters;\\Characters of the term
}
```

In a visual manner, the index for the term '上呼吸道感染{*upper respiratory tract infection*, *shang*4 *hu*1 *xi*1 *dao*4 *gan*3 *ran*3}' is indexed as follows:

[4] http://ictclas.org/

[5] http://pinyin4j.sourceforge.net/

```
INDEX = { \\ '上呼吸道感染'
    [上呼吸道 感染] \\Words
    [上 感] \\WordInitials
    [shang4 hu1 xi1 dao4 gan3 ran3] \\Pinyins
    [s h x d g r];\\ PinyinInitials
    [ang u i ao an an] \\ PinyinFinals
    [上 呼 吸 道 感 染] \\Characters
}
```

Note in the INDEX structure that *Word initials*, *Pinyin initials*, *Pinyin finals* and *non-Chinese characters* are involved. They are considered as fields in the index due to the following reasons:

- *Word initials*: Some informal terms are created using abbreviation of word initials, e.g., $E1$ in Table 1.
- *Pinyin initials*: Some informal terms are created using abbreviation of pinyin initials, e.g., $E2$ in Table 1.
- *Pinyin finals*: Some informal terms are created due to an error input, e.g., '阿司品林(*a*1 *si*1 *pin*3 *lin*2)' which corresponds to '阿司匹林(*asprin, a*1 *si*1 *pi*1 *lin*2)'.

In this work, all the medical terms in the Chinese medical ontology are dumped into the Lucene system.

3.3 Term Retrieval and Ranking

As standard medical terms are indexed in multiple fields, the informal term input to the Lucene system should also be pre-processed in the same way so as to match the standard terms in the according fields. As showed in Fig 1, the words, characters and pinin's are first used to form a multi-filed query. Using the informal term '上感(*shang*4 *gan*3)' in example $E1$ as example, the query is formed as follows:

```
INDEX = { \\ '上感'
    Words = [上 感]
    WordInitials = [上 感]
    Pinyins = [shang4 gan3]
    PinyinInitials = [s g]
    PinyinFinals = [ang an]
    Characters = [上 感]
}
```

The $BM25$ algorithm[6] is a standard relevance calculation algorithm. Note that the $BM25$ algorithm assigns the multiple fields equal weights. However, our empirical study indicates that the weights vary significantly considering their

[6] http://en.wikipedia.org/wiki/Okapi_BM25

Table 2. Empirical weights of the multiple fields in the index

Field	Weight
Words	1
WordInitials	0.1
Pinyins	3
PinyinInitials	0.1
PinyinFinals	2
Characters	1

contribution to term retrieval. Using 100 human judged pairs of informal terms and their standard counterparts, we obtain the weights for the fields in Table 2.

At last, the $BM25$ algorithm output top N standard terms which hold biggest relevance score.

4 Evaluation

4.1 Setup

Dataset
In this work, 300 pairs of informal medical terms and their standard counterparts are compiled by medical experts. The dataset covers 125 Chinese abbreviations, 48 pinyin abbreviations and 127 typos.

In our medical ontology, 48,000 medical terms are covered, which are used as standard terms.

Evaluation Metric
We first adopt precision in top N terms (i.e. p@N) as evaluation metric. That is, we calculate percentage of correctly normalized terms amongst all the input informal terms. Meanwhile, we adopt execution time to compare computational complexity of the methods.

4.2 Experiment 1: Normalization Methods

In this experiment, we intend to compare our proposed method against the following two baseline methods:

- *Edit distance* (EDDis): Similarity of the input informal terms and standard term is calculated using edit distance[7] in the six fields in Section 3.3. The wights in Table 2 are used to combine the six similarity values in a linear manner, thus an overall similarity value is obtained for the input term and the standard term.
- *Multi-filed cosine similarity* (MSim): This method differs from the edit distance method in calculating similarity with cosine formula.

[7] http://en.wikipedia.org/wiki/Edit_distance

In the proposed IR-based normalization method (IRNorm), the retrieval and ranking process are delivered with $BM25$ within Lucene. Experimental results are presented in Table 3.

Table 3. Experimental of different methods for medical term normalization

Method	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*

* The total time for indexing the 48,000 standard terms is 30 seconds.

Discussion

It can be seen from Table 3 that the precision values of MSim and IRNorm are close, while are both higher than EDDis in both top 5 and top 10 terms. This indicate that cosine distance and $BM25$ are both more effective than edit distance in term normalization.

We also notice that EDDis and MSim require much longer computing time. Counting in the indexing time, i.e. 1620 seconds in total, the proposed IRNorm method is much faster than EDDis and MSim. This ascribes to the Lucene engine.

4.3 Experiment 2: The Fields in the Index

In this experiment, we seek to evaluate contribution of the fields in the index. Using different configuration, we develop various implementations of the proposed method, showed in Table 4.

Table 4. Implementations of our method with different configuration

Method ID	Word	Character	Pinyin
IRNorm-A	Y	N	N
IRNorm-B	Y	Y	N
IRNorm-C	Y	N	Y
IRNorm-D*	Y	Y	Y

* This is the method used in Experiment 1.

Experimental results are presented in Table 5.

Discussion

It can be seen from Table 5 that IRNorm-B outperforms IRNorm-A by 0.226 on p@5 and by 0.241 on p@10. This indicates that the Chinese character makes significant contribution to term normalization. Comparing IRNorm-A and IRNorm-C, we find pinyin makes more contribution, i.e., improving by 0.297 on p@5 and by 0.311 on p@10. When the Chinese character and pinyin are both

Table 5. Experimental results (p@N) of different implementations of our method

Implementation	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

used, the outperformance is significant. This indicates that the fields that are defined in this work are indeed helpful in discovering the standard counterparts of the input informal terms.

5 Conclusion and Future Work

The informal medical terms pose huge challenge to medical business intelligence systems. In this work, the term normalization task is accomplished with an information retrieval framework using multi-field index. The contributions of this work are summarized as follows. Firstly, the proposed method considers words, Chinese characters and pinyin's in standard term matching, which makes term normalization more accurate. Secondly, with the multi-field index under information retrieval framework, the normalization process is made much faster. Experiments show the proposed method is very effective.

Note that this work is still preliminary. The following work is planned. First, only intra-term features are considered in term normalization. Therefore, we will explore how context helps to normalize the informal medical terms. Secondly, in our term normalization system, the input is informal term, which is detected by human experts. However, in many cases the informal term should be detected automatically by machines. Thus in the future we will develop the informal term detection algorithm.

Acknowledgement. This work is supported by National Science Foundation of China (NSFC: 61272233). We thank the anonymous reviewers for the valuable comments.

References

1. Koh, H., Tan, G.: Data mining applications in healthcare. J. Healthcare Inf. Manag. 19(2), 64–72 (2005)
2. Suominen, H., et al.: Overview of the shARe/CLEF eHealth evaluation lab 2013. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 212–231. Springer, Heidelberg (2013)

3. Campbell, K.E., Oliver, D.E., Shortliffe, E.H.: The unified medical language system: Toward a collaborative approach for solving terminologic problems. JAMIA 5(1), 12–16 (1998)
4. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research 32(database issue), 267–270 (2004)
5. Kim, M.Y., Goebel, R.: Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking. In: 2010 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB), pp. 1–5. IEEE (2010)
6. Wu, Y., Denny, J., Rosenbloom, S., Miller, R., Giuse, D., Xu, H.: A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In: AMIA Annu. Symp., 997–1003 (2012)
7. Sproat, R., Black, A.W., Chen, S.F., Kumar, S., Ostendorf, M., Richards, C.: Normalization of non-standard words. Computer Speech & Language 15(3), 287–333 (2001)
8. Xia, Y., Wong, K.F., Li, W.: A phonetic-based approach to chinese chat text normalization. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 993–1000. Association for Computational Linguistics, Stroudsburg (2006)