Medical Record Retrieval and Extraction for Professional Information Access

Chia-Chun Lee, Hen-Hsen Huang, and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
{cclee, hhhuang}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

Abstract

This paper analyzes the linguistic phenomena in medical records in different departments, including average record size, vocabulary, entropy of medical languages, grammaticality, and so on. Five retrieval models with six pre-processing strategies on different parts of medical records are explored on an NTUH medical record dataset. Both coarse-grained relevance evaluation on department level and fine-grained relevance evaluation on course and treatment level are conducted. Query accesses to the medical records in medical languages of smaller entropy tend to have better performance. The departments related to generic parts of body such as Departments of Internal Medicine and Surgery may confuse the retrieval, in particular, for Departments of Oncology and Neurology. Okapi model with stemming achieves the best performance on both department and course and treatment levels.

1 Introduction

Mining wisdom of crowds from heterogeneous domains to support various applications has attracted much attention in this decade. The contributors of knowledge may be various from common users to domain experts. Internet forums, weblogs, and wikis are contributed by general users, while scientific documents and medical records are written by experts. Literature mining aims at extracting knowledge from biomedical literature, constructing a knowledge base (semi-)automatically, and finding new knowledge [Jen06]. Comparatively, medical text mining aims at discovering medical knowledge from electronic patient records. There are many potential applications, e.g., comorbidities and disease correlations [Got12], acute myocardial infarction mining [Hei01], assessment of healthcare utilization and treatments [Ram11], outpatient department recommendation [Hua12], virtual patient in health care education, and so on.

Finding relevant information is the first step to mining knowledge from diverse sources. Different information retrieval systems have been developed to meet these needs. This paper focuses on professional information access and addresses the supports for experts of medical domain. PubMed, which comprises more than 22 million citations for biomedical literature from MEDLINE, provides information retrieval engines for finding biomedical documents. Information retrieval on medical records has been introduced to improve healthcare services [Her09][Hua12]. Medical records are similar to scientific documents in that both are written by domain experts, but they are different from several aspects such as authorship, genre, structure, grammaticality, source, and privacy. Biomedical literatures are research findings of researchers. The layout of a scientific paper published in journals and conference proceedings are often composed of problem specification, solutions, experimental setup, results, discussion and conclusion. To gain more impacts, scientific literatures are often made available to the public. Grammatical correctness and readability are the basic requirements for publication.

In contrast, medical records are patients' treatments by physicians when patients visit hospitals. The basic layout consists of a chief complaint, a brief history, and a course and treatment. From the ethical and legal aspects, medical records are privacy-sensitive. Release of medical records is restricted by government laws. Medical records are frequently below par in grammaticality. That is not a problem for the understanding by physicians, but is an issue for retrieval.

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

Case study is indispensable for learning medical knowledge. The course and treatments of similar cases provide important references, in particular, for medical students or junior physicians. How to retrieve relevant medical records effectively and efficiently is an essential research topic. TREC 2011 [Voo11] and 2012 [Voo12] Medical

Records track provides test collections for patient retrieval based on a set of clinical criteria. Several approaches such as concept-based [Koo11], query expansion [Din11], and knowledge-based [Dem11] have been proposed to improve the retrieval performance. In this paper, we investigate medical record retrieval on an NTUH dataset provided by National Taiwan University Hospital. Given a chief complaint and/or a brief history, we would like to find the related medical records, and propose examination, medicine and surgery that may be performed for the input case.

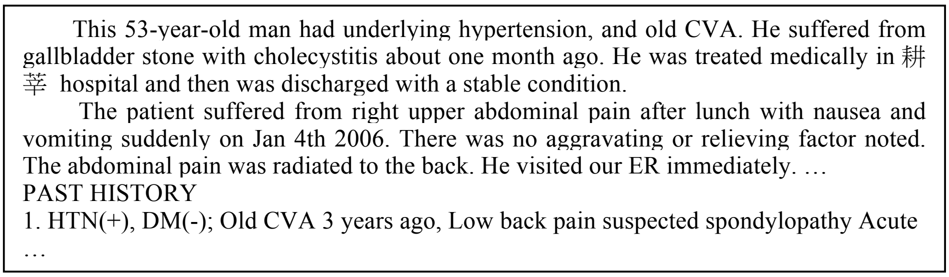
The structure of this paper is organized as follows. The characteristics of the domain-specific dataset are addressed and analyzed in Section 2. Several information retrieval models and medical term extraction methods are explored on the dataset in Section 3. Both coarse-grained relevance evaluation on department level and fine-grained relevance evaluation on course and treatment level are conducted and discussed in Section 4. Finally, Section 5 concludes the remarks.

2 Description of the NTUH Medical Record Dataset

In the NTUH dataset, almost all medical records are written in English. A medical record is composed of three major parts, including a chief complaint, a brief history, and a course and treatment. A chief complaint is a short statement specifying the purpose of a patient's visit and the patient's physical discomfort, e.g., *Epigastralgia for 10 days*, *Tarry stool twice since last night*, and so on. It describes the symptoms found by the patient and the duration of these symptoms. A brief history summarizes the personal information, the physical conditions, and the past medical treatment of the patient. In an example shown in Figure 1, the first paragraph lists the personal information and the physical conditions, and the second paragraph shows the past medical treatment. A course and treatment describes the treatment processes and the treatment outcomes in detail. Figure 2 is an example of a course and treatment, where medication administration, inspection, and surgery are enclosed in <a></a>, <i></i>, and <s></s> pairs, respectively.

There are 113,625 medical records in the NTUH experimental dataset after those records consisting of scheduled cases, empty complaints, complaints written in Chinese, and treatments without mentioning any examination, medicine, and surgery are removed. Table 1 lists mean (μ) and standard deviation (σ) of chief complaint (CC), brief history (BH), course and treatment (CT), and medical record (MR) in terms of the number of words used in the corresponding part. Here a word is defined to be a character string separated by spaces. The patient and the physician names are removed from the dataset for the privacy issues. The brief history is the longest, while the chief complaint is the shortest.

The 113,625 medical records are categorized into 14 departments based on patients' visits. The statistics is illustrated in Table 2. Departments of Internal Medicine and Surgery have the first and the second largest amount of data, while Departments of Dental and Dermatology have the smallest amount. Table 3 shows the length distribution of these 14 departments. For the chief complaint, Department of Urology has the smallest mean, and Department of Dermatology has the largest mean. For the brief history, Department of Ophthalmology has the smallest mean and standard deviation, and Department of Psychiatry has the largest mean. Overall, Department of Dental has the smallest mean, and Department of Psychiatry has the largest mean as well as standard deviation.



This 53-year-old man had underlying hypertension, and old CVA. He suffered from gallbladder stone with cholecystitis about one month ago. He was treated medically in 耕莘 hospital and then was discharged with a stable condition.

The patient suffered from right upper abdominal pain after lunch with nausea and vomiting suddenly on Jan 4th 2006. There was no aggravating or relieving factor noted. The abdominal pain was radiated to the back. He visited our ER immediately. …

PAST HISTORY

1. HTN(+), DM(-); Old CVA 3 years ago, Low back pain suspected spondylopathy Acute …

Figure 1: A Brief History

After admission, <a> Heparin </a> was given immediately. Venous duplex showed left common iliac vein partial stenosis. Pelvic-lower extremity revealed bilateral mid. femoral vein occlusion. <i> Angiography </i> showed total occlusion of left iliac vein, femoral vein and popliteal vein. IVC filter was implanted. Transcatheter intravenous urokinase therapy was started on 1/11 for 24 hours infusion. Follow up <i> angiography </i> showed partial recanalization of left iliac vein. Stenting was donefrom distal IVC through left common iliac vein to external iliac vein. <s> Ballooming </s> was also performed. …

Figure 2: A Course and Treatment

Table 1: Mean and Standard Deviation of Medical Records in Words

component	mean (μ)	standard deviation (σ)
chief complaint (CC)	7.88	3.75
brief history (BH)	233.46	163.69
course and treatment (CT)	110.28	145.04
medical record (MR)	351.62	248.51

Table 2: Distribution of the Medical Records w.r.t. Department Type

Dental	1,253	Dermatology	1,258	Ear, Nose & Throat	7,680
Internal Medicine	**34,396**	Neurology	2,739	Obstetrics & Gynecology	5,679
Oncology	4,226	Ophthalmology	3,400	Orthopedics	8,814
Pediatrics	11,468	Rehabilitation	1,935	Psychiatry	1,656
Surgery	**23,303**	Urology	5,818		

Table 3: Mean and Standard Deviation of Medical Records in Each Department Type

	μ	σ	μ	σ	μ	σ
	Dental		Dermatology		Ear, Nose & Throat	
CC	9.14	4.13	11.25	3.93	7.17	2.3
BH	138.97	65.45	232.4	108.58	158.33	84.81
CT	23.31	35.93	123.71	140.3	47.46	27.31
MR	171.41	87.27	367.35	197.93	212.95	95.48
	Internal Medicine		Neurology		Obstetrics & Gynecology	
CC	7.8	4.75	10.17	3.64	7.8	2.69
BH	278.72	154.61	251.87	127.61	175.13	129.89
CT	162.28	182.69	141.87	115.52	53.45	55.53
MR	448.8	257.04	403.91	190.02	236.38	156.11
	Oncology		Ophthalmology		Orthopedics	
CC	8.29	3.34	8.21	2.44	8.42	3.73
BH	418.46	201.19	117.93	47.73	131.96	70.26
CT	170.44	193.36	49.59	32.04	44.75	38.0
MR	597.19	301.34	175.73	65.87	185.14	88.71
	Pediatrics		Rehabilitation		Psychiatry	
CC	7.52	2.84	9.27	2.82	10.01	4.79
BH	293.46	189.83	346.09	186.26	521.73	287.7
CT	137.77	184.69	183.4	101.47	162.44	96.51
MR	438.75	291.07	538.77	227.96	694.19	320.3
	Surgery		Urology			
CC	7.73	3.04	6.26	2.78		
BH	191.03	126.37	152.96	121.31		
CT	84.22	103.71	44.33	59.26		
MR	282.98	179.89	203.54	148.27		

From the linguistic point of view, we also investigate the vocabulary size and entropy of the medical language overall for the dataset and individually for each department. Table 4 summarizes the statistics. Shannon [Sha51] estimated word entropy for English as 11.82 bits per word, but there has been some debate about this estimate, with Grignetti [Gri64] estimating entropy to be 9.8 bits per word. In the NTUH medical dataset, the entropy is 11.15 bits per word, a little smaller than Shannon entropy and larger than Grignetti entropy. Departments related to definite parts of body, e.g., dental, ear, nose & throat, ophthalmology and orthopedics, have lower entropy.

Comparatively, departments related to generic parts have larger entropy. In particular, Department of Ophthalmology has the lowest entropy, while Department of Internal Medicine has the largest entropy. Medical records are frequently below par in grammaticality. Spelling errors are very common in this dataset. Some common erroneous words and their correct forms enclosed in parentheses are listed below for reference: histropy (history), ag (ago/age), withour (without), denid (denied), and recieved (received). Some words are ambiguous in the erroneous form, e.g., "ag" can be interpreted as "ago" or "age" depending on its context. Besides grammatical problems, shorthand notation or abbreviation occurs very often. For example, "opd" is an abbreviation of "outpatient department" and "yrs" is a shorthand notation of "years-old". Furthermore, physicians tend to mix English and Chinese in the NTUH dataset. In Department of Psychiatry, the chief complaint of psychiatric disorder patients is more descriptive, and it is hard to write down the descriptions completely in English. Physicians express the patients' descriptions bilingually, e.g., "Chronic insomnia for 10+ years 吃了 10 顆 FM2 還是睡不著, 希望住院調整睡." Furthermore, physicians tend to name hospitals in Chinese.

Table 4: Vocabulary Size and Entropy of the Medical Language w.r.t. Department Type

Vocabulary Size	Entropy	Vocabulary Size	Entropy	Vocabulary Size	Entropy
Dental		Dermatology		Ear, Nose & Throat	
15,036	**9.74**	26,914	10.32	48,452	**9.88**
Internal Medicine		Neurology		Obstetrics & Gynecology	
415,279	11.06	55,301	10.62	65,760	10.46
Oncology		Ophthalmology		Orthopedics	
101,361	10.81	27,765	**9.70**	47,082	**9.79**
Pediatrics		Rehabilitation		Psychiatry	
175,555	10.86	51,328	10.50	67,390	10.64
Surgery		Urology		Overall	
203,677	10.76	53,853	10.25	786,666	11.15

3 Retrieval and Extraction Models

The retrieval scenario is specified as follows. Given a chief complaint and/or a brief history, physicians plan to retrieve the similar cases from the historical medical records and reference to the possible course and treatments. Chief complaints and/or brief histories in the historical medical records can be regarded as queries. Words may be stemmed and stop words may be removed before indexing. Spelling checker is introduced to deal with grammaticality issue. Besides words, medical terms are also recognized as indices. Different IR models can be explored on different parts of medical records. In the empirical study, Lemur Toolkit (http://www.lemurproject.org/) is adopted and five retrieval models including tf-idf, okapi, KL-divergence, cosine, and indri are experimented.

The medical terms such as examination, medicine, and surgery are extracted from the course and treatment of the retrieved medical records. Medical term recognition [Aba11] is required. Ontology-based and pattern-based approaches are adopted. The ontology-based approach adopts the resources from the Unified Medical Language System (UMLS) maintained by National Library of Medicine. The UMLS covers a wide range of terms in medical domain, and relations between these medical terms. Among these resources, the Metathesaurus organizes medical terms into groups of concepts. Moreover, each concept is assigned at least one Semantic Type. Semantic Types provide categorization of concepts at a more general level, and therefore are well-suited to be incorporated. The pattern-based approach adopts patterns such as "**SURGERY** was performed on **DATE**" to extract medical terms. The idea comes from the special written styles of medical records. A number of patterns frequently repeat in medical records. The following lists some examples for the pattern "**SURGERY** was performed on **DATE**": **paracentesis** was performed on **2010-01-08**, **repositioning** was performed on **2008/04/03**, **incision and drainage** was performed on **2010-01-15**, and **tracheostomy** was performed on **2010/1/11**.

We follow our previous work [Che12] to extract frequent patterns from medical record dataset and apply them to recognize medical terms. The overall schedule is summarized as follows.

(a) Medical Entity Classification: Recognize medical named entities including surgeries, diseases, drugs, etc. by the ontology-based approach, transform them into the corresponding medical classes, and derive a new corpus.

(b) Frequent Pattern Extraction: Employ n-gram models in the new corpus to extract a set of frequent patterns.

(c) Linguistic Pattern Extraction: For each pattern, randomly sample sentences having this pattern, parse these sentences, and keep the pattern if there is at least one parsing sub-tree for it.

(d) Pattern Coverage Finding: Check coverage relations among higher order patterns and lower order patterns, and remove those lower patterns being covered.

To evaluate the performance of the retrieval and extraction models, 10-fold cross validation is adopted. We conduct two-phase evaluation. In the first phase, the input query is a chief complaint and the output is the retrieved top-*n* medical records. We aim to evaluate the quality of the returned *n* medical records. There is no ground truth or relevance judgments available, surrogate relevance judgments are therefore used. Recall that each medical record belongs to a department. Let the input chief complaint belong to department *d*, and the departments of the top-*n* retrieved medical records be d_1, d_2, …, d_n. Here, we postulate that medical record *i* is relevant to the input chief complaint, if d_i of medical record *i* is equal to *d*. In this way, we can compute precision@k, mean average precision (MAP), and nDCG as traditional IR. In addition, we can regard the returned *n* medical records as a cluster and compute the department distribution of the cluster. The retrieval is regarded as correct if the dominant department of the cluster is the same as the department of the input query (i.e., the input chief complaint). In this way, we can compute the confusion matrix among actual and proposed departments and observe the effects on retrieval performance.

In the second phase, we conduct much finer evaluation. The input is a chief complaint and a brief history, and the output is top-1 course and treatment selected from the historical NTUH medical records. Recall that examination, medicine and surgery are three key types of medical entities specified in a course and treatment. We would like to know if the retrieved medical record adopts the similar course and treatment as the input query. Thus the evaluation unit is the three types of entities. We extract examinations, medicines and surgeries from the courses and treatments of an input query and the retrieved medical record, respectively, by medical term recognition. They are named as *GE*, *GM*, and *GS* for ground truth (i.e., the course and treatment of the input query), and *PE*, *PM*, and *PS* for the proposed treatment (i.e., the course and treatment of the returned medical record), respectively. The Jaccard's coefficient between the ground truth and the proposed treatment is a metric indicating if the returned medical records are relevant and interesting to physicians. It is defined as: total number of common entities in the ground truth and the proposed answer divided by sum of the entities in the ground truth and the proposed answer for each query. The evaluation is done for each medical entity type. That is, Jaccard's coefficient for examination=|GE∩PE|/|GE∪PE|, Jaccard's coefficient for medicine=|GM∩PM|/|GM∪PM|, and Jaccard's coefficient for surgery=|GS∩PS|/|GS∪PS|. Note that the denominator will be zero, if both the ground truth and the proposed answer do not contain any medical entities of the designated type. In this case, we set Jaccard's coefficient to be 1. The average of the Jaccard's coefficients of all the input queries is considered as a metric to evaluate the performance of the retrieval model on the treatment level.

Table 5: MAP and nDCG of Retrieval Models on the Department Level with Different Strategies

model	metric	S1	S2	S3	S4	S5	S6
Top 5							
tf-idf	MAP	0.6858	0.6776	0.6860	0.6780	0.6700	0.6685
	nDCG	0.7529	0.7456	0.7535	0.7461	0.7385	0.7370
okapi	MAP	**0.6954**	**0.6871**	**0.6965**	**0.6875**	**0.6800**	**0.6774**
	nDCG	**0.7622**	**0.7545**	**0.7626**	**0.7551**	**0.7489**	**0.7469**
kl	MAP	0.6715	0.6634	0.6692	0.6612	0.6691	0.6654
	nDCG	0.7396	0.7316	0.7385	0.7305	0.7380	0.7350
cos	MAP	0.6857	0.6818	0.6868	0.6827	0.6521	0.6503
	nDCG	0.7520	0.7485	0.7534	0.7488	0.7217	0.7203
indri	MAP	0.6638	0.6582	0.6604	0.6558	0.6557	0.6527
	nDCG	0.7328	0.7274	0.7305	0.7264	0.7251	0.7220
		S1	S2	S3	S4	S5	S6
Top 10							
tf-idf	MAP	0.6651	0.6584	0.6660	0.6590	0.6502	0.6487
	nDCG	0.7481	0.7420	0.7486	0.7422	0.7348	0.7330
okapi	MAP	**0.6734**	**0.6672**	**0.6749**	**0.6678**	**0.6588**	**0.6566**
	nDCG	**0.7559**	**0.7498**	**0.7564**	**0.7498**	**0.7427**	**0.7404**
kl	MAP	0.6517	0.6444	0.6499	0.6430	0.6489	0.6465
	nDCG	0.7362	0.7297	0.7352	0.7285	0.7329	0.7307
cos	MAP	0.6648	0.6611	0.6660	0.6622	0.6340	0.6331
	nDCG	0.7473	0.7437	0.7481	0.7447	0.7186	0.7181
indri	MAP	0.6446	0.6395	0.6422	0.6380	0.6365	0.6339
	nDCG	0.7305	0.7256	0.7285	0.7246	0.7221	0.7192

4 Results and Discussion

Table 5 shows the coarse-grained relevance evaluation on department level. Five retrieval models shown in the 1st column with six strategies (S1)-(S6) are explored. These six strategies are defined as follows. Top 5 and top 10 medical records are retrieved and compared. For strategies S5 and S6, we extract gender (male/female), age (0-15, 16-45, 46-60, 61+), and other information from brief history besides chief complaints.

S1: using chief complaints
S2: S1 with stop word removal
S3: S1 with porter stemming
S4: S1 with both stop word removal and porter stemming
S5: using chief complaints and the first two sentences in brief histories
S6: S5 with porter stemming

Overall, the performance tendency is okapi>tf-idf>cos>kl>indri no matter which strategies are used. Removing stop words tend to decrease the performance. Using porter stemming is useful when chief complaints are employed only. Introducing brief histories decreases the performance. Okapi retrieval model with strategy S3 performs the best when top 5 medical records are retrieved. In fact, Okapi+S3 is not significantly better than Okapi+S1, but both are significantly better than Okapi with other strategies (*p* value <0.0001) on MAP and nDCG. When S3 is adopted, Okapi is significantly better than the others.

We further evaluate the retrieval models with precision@k shown in Table 6. The five retrieval models at the setting *k*=1 are significantly better than those at *k*=3 and *k*=5. Most of the precision@k are larger than 0.7 at *k*=1. It means the first medical record retrieved is often relevant. Okapi with strategy S3 is still the best under precision@k. Moreover, we examine the effects of the parameter *n* in the medical record retrieval. Only the best two retrieval models in the above experiments, i.e., tf-idf and okapi with strategy S3, are shown in Figure 3. We can find MAP decreases when *n* becomes larger in both models. It means noise is introduced when more medical records are reported. The Okapi+S3 model is better than the tf-idf+S3 model in all the settings.

Table 6: precision@k of Retrieval Models on the Department Level with Different Strategies

model	precision@k	S1	S2	S3	S4	S5	S6
tf-idf	*k*=1	0.7185	0.7103	0.7188	0.7105	0.7031	0.7013
okapi		**0.7280**	**0.7197**	**0.7293**	**0.7203**	**0.7136**	**0.7109**
kl		0.7041	0.6958	0.7020	0.6933	0.7021	0.6984
cos		0.7184	0.7138	0.7193	0.7149	0.6857	0.6827
indri		0.6960	0.6907	0.6926	0.6879	0.6880	0.6857
tf-idf	*k*=3	0.6259	0.6196	0.6269	0.6204	0.6132	0.6117
okapi		**0.6371**	**0.6316**	**0.6384**	**0.6326**	**0.6238**	**0.6231**
kl		0.6073	0.5997	0.6055	0.5988	0.6120	0.6105
cos		0.6273	0.6236	0.6279	0.6245	0.5983	0.5970
indri		0.5986	0.5947	0.5967	0.5935	0.5986	0.5973
tf-idf	*k*=5	0.5963	0.5911	0.5980	0.5928	0.5863	0.586
okapi		**0.6072**	**0.6034**	**0.6099**	**0.605**	**0.5973**	**0.5965**
kl		0.5775	0.5719	0.5770	0.5725	0.5842	0.5838
cos		0.5972	0.5933	0.5984	0.5951	0.5741	0.5741
indri		0.5698	0.5670	0.5691	0.5676	0.5713	0.5702

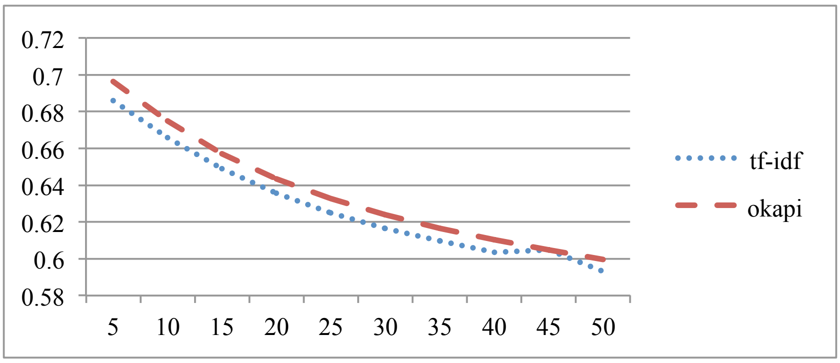


Figure 3: MAPs of tf-idf and okapi under Different *n*'s

Table 7 further shows the retrieval performance in terms of MAP, nDCG and precision@k with respect to department type. Note four departments have entropy less than 10 shown in Table 4, i.e., Departments of Dental, Ear, Nose & Throat, Ophthalmology, and Orthopedics. The performances of query accesses to medical records in these departments are more than 0.8200 in all the metrics. In particular, the retrieval performances for Department of Ophthalmology are even more than 0.9155. Comparatively, Department of Internal Medicine, which has the largest entropy, achieves the average performance. Department of Oncology gets the worst retrieval performance because tumor may occur in different organs. The precision@1 to access medical records in this department is only 0.3685, which is the worst of all.

Table 8 lists the confusion matrix among department types. The diagonals show how many percentages the dominant department in the retrieved medical record cluster is the same as the actual department. Larger diagonal values mean good retrieval performance. The results are quite similar to those in Table 7. The values of Dental-Dental, Ear&Nose&Throat-Ear&Nose&Throat, Ophthalmology-Ophthalmology and Orthopedics-Orthopedics are larger than those of other department pairs in the corresponding rows. In contrast, the value of Oncology-Oncology is 0.1545, which is even smaller than the values of Oncology-Internal Medicine (i.e., 0.4792) and Oncology-Surgery (i.e., 0.1805). That may be because tumor is often found in Department of Internal Medicine, and treated in Department of Surgery. Similarly, the access related to Department of Neurology is also worse in Table 7. Table 8 shows the value of Neurology-Internal Medicine (i.e., 0.2950) is very close to that of Neurology-Neurology (i.e., 0.3552).

Table 7: Retrieval Performance w.r.t. Department Type Using Okapi Retrieval Model and Strategy S3

Department	MAP@5	nDCG@5	MAP@10	nDCG@10	precision@1
Dental	**0.8545**	**0.8825**	**0.8295**	**0.8744**	**0.8755**
Dermatology	0.6531	0.7083	0.6263	0.7003	0.6901
Ear, Nose & Throat	**0.8443**	**0.8770**	**0.8282**	**0.8715**	**0.8640**
Internal Medicine	0.7001	0.7867	0.6695	0.7688	0.7381
Neurology	0.4843	0.5762	0.4612	0.5731	0.5232
Obstetrics & Gynecology	0.7779	0.8121	0.7635	0.8100	0.8000
Oncology	0.3233	0.3847	0.3236	0.4185	0.3685
Ophthalmology	**0.9265**	**0.9419**	**0.9155**	**0.9371**	**0.9377**
Orthopedics	**0.8518**	**0.8888**	**0.8326**	**0.8802**	**0.8736**
Pediatrics	0.6667	0.7278	0.6509	0.7290	0.6977
Rehabilitation	0.6088	0.6772	0.5921	0.6771	0.6390
Psychiatry	**0.8323**	**0.8631**	**0.8183**	**0.8608**	**0.8487**
Surgery	0.6120	0.6971	0.5889	0.6943	0.6535
Urology	0.7651	0.8035	0.7494	0.8037	0.7873

Table 8: Confusion Matrix among Departments

Actual Dept.	Dominant Department (%)														
	Dent	Derm	ENT	Med	Neur	O&G	Onc	Ophth	Ortho	Pedi	Reha	Psyc	Surg	Urol	Un
Dent	**76.94**	0.32	6.62	4.87	0.08	0.40	0.72	0.00	1.44	0.56	0.00	0.00	7.74	0.32	0.00
Derm	0.32	54.45	2.70	22.97	0.79	0.32	0.40	0.32	2.62	2.62	0.08	0.16	11.76	0.48	0.00
ENT	0.56	0.09	**83.62**	6.80	0.27	0.18	0.87	0.35	0.38	1.00	0.01	0.03	5.64	0.18	0.01
Med	0.09	0.40	1.14	74.39	0.82	0.73	1.08	0.37	1.50	5.14	0.96	0.24	11.66	1.48	0.00
Neur	0.04	0.22	1.10	29.50	35.52	0.29	0.73	1.64	2.08	2.63	11.46	1.57	12.38	0.84	0.00
O&G	0.18	0.05	0.67	11.92	0.07	70.73	0.35	0.11	0.83	1.34	0.19	0.02	10.02	3.52	0.00
Onc	0.35	0.21	6.25	47.92	0.62	1.37	15.45	0.33	4.54	2.70	0.50	0.09	18.05	1.61	0.00
Ophth	0.03	0.06	0.85	3.18	0.06	0.12	0.09	**90.91**	0.71	0.47	0.03	0.00	3.12	0.38	0.00
Ortho	0.19	0.16	0.68	3.03	0.31	0.15	0.35	0.12	**85.28**	0.27	0.16	0.03	8.91	0.34	0.01
Pedi	0.17	0.33	1.13	29.30	0.43	0.58	0.35	0.22	0.81	58.07	0.18	0.19	7.39	0.82	0.01
Reha	0.05	0.05	0.47	15.30	8.37	0.21	0.31	0.21	5.37	0.41	57.52	0.05	9.66	2.02	0.00
Psyc	0.00	0.06	0.60	11.78	0.79	0.06	0.60	0.18	0.36	1.57	0.12	**80.01**	3.74	0.12	0.00
Surg	0.27	0.25	2.73	30.70	1.10	0.97	0.79	0.64	5.18	3.51	0.74	0.11	51.25	1.76	0.00
Urol	0.09	0.03	0.81	12.98	0.07	1.58	0.52	0.14	1.01	1.39	0.22	0.03	10.06	71.04	0.03

Table 9: Jaccard's Coefficients of Retrieval Models on the Course and Treatment Level with Different Strategies

	Top-1	S1	S2	S3	S4	S5	S6
tf-idf	examination	0.3332	0.3109	0.3515	0.3289	0.3728	0.3727
	medicine	0.2501	0.2445	0.2589	0.2539	0.3166	0.3147
	surgery	0.1115	0.1154	0.1131	0.1168	0.1851	0.1835
okapi	examination	0.3448	0.3376	0.3499	0.3447	**0.3816**	0.3810
	medicine	0.2995	0.2980	0.3000	0.2988	0.3289	0.3278
	surgery	0.1406	0.1397	0.1394	0.1406	**0.1954**	0.1936
kl	examination	0.4351	0.4017	0.4399	0.4076	0.3690	0.3679
	medicine	0.2222	0.2370	0.2245	0.2389	0.3112	0.3101
	surgery	0.0847	0.0961	0.0844	0.0950	0.1821	0.1803
cos	examination	0.3362	0.3305	0.3437	0.3362	0.3814	0.3826
	medicine	0.2846	0.2865	0.2897	0.2905	0.3292	0.3291
	surgery	0.1358	0.1393	0.1339	0.1376	0.1882	0.1875
indri	examination	0.4501	0.4202	**0.4535**	0.4259	0.3639	0.3636
	medicine	0.2035	0.2257	0.2055	0.2267	0.3042	0.3035
	surgery	0.0776	0.0898	0.0764	0.0879	0.1758	0.1743

Table 9 lists the fine-grained relevance evaluation on the course and treatment level with Jaccard's coefficient. Total 663 examinations, 2,165 medicines, and 1,483 surgeries are used in the treatments. Total 54,679, 64,607, and 88,647 medical records mention examinations, medicines, and surgeries in their treatments. We count the number of the same examinations (medicines or surgeries) appearing in both ground truth and the treatment of the top-1 returned medical record. The number is normalized by total number of examinations (medicines or surgeries) in both treatments for each query. If both do not recommend any examinations (medicines or surgeries), the Jaccard's coefficient is regarded as 1. The five retrieval models and the six strategies used in the above experiments are explored again in the fine-grained evaluation. Overall, the performance of examination prediction is larger than that of medicine prediction, which is larger than that of surgery prediction. Considering brief history (i.e., strategies S5 and S6) benefits medicine and surgery prediction. The experimental results show that Okapi model with strategy S5 achieves the best performance on medicine and surgery prediction (i.e., 0.3289 and 0.1954), and Indri with strategy S3 achieves the best performance on examination prediction (i.e., 0.4535).

5 Conclusion

This paper studies the medical record retrieval and extraction with different retrieval models under different strategies on department and course and treatment levels. Both coarse-grained and fine-grained relevance evaluations with various metrics are conducted. The medical records in medical languages of smaller entropy tend to have better retrieval performance. The departments related to generic parts of body such as Departments of Internal Medicine and Surgery may confuse the retrieval, in particular, for Departments of Oncology and Neurology. Okapi model achieves the best on department and treatment levels (in particular, medicine prediction and surgery prediction). To construct an evaluation dataset for medical record retrieval and extraction is challenging because the assessors which are domain experts cost much. In this paper, we postulate that the medical records belong to the same departments as the input queries are relevant. Such an evaluation may be underestimated because cross department is not necessarily wrong in real cases. For example, the treatment of tumors may be related to more than one department. Real user study is necessary for advanced evaluation. Besides, medical records may be in more than one language. Cross language medical retrieval will be explored in the future.

Acknowledgments

Research of this paper was partially supported by National Science Council (Taiwan) under the contract NSC 101-2221-E-002-195-MY3. We are very thankful to National Taiwan University Hospital for providing NTUH the medical record dataset. The authors also thank anonymous reviewers for their helpful comments.

References

[Aba11] A. B. Abacha, and P. Zweigenbaum. Medical entity recognition: a comparison of semantic and statistical methods. Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, 2011, 56-64.

[Che12] H.-B. Chen, H.-H. Huang, J. Tjiu, C.-T. Tan and H.-H. Chen. A statistical medical summary translation system. *Proceedings of 2012 ACM SIGHIT International Health Informatics Symposium*, January 2012, 101-110.

[Dem11] D. Demner-Fushman, S. Abhyankar, et al. A Knowledge-Based Approach to Medical Records Retrieval. Proceedings of TREC, 2011.

[Din11] D. Dinh and L. Tamine. IRIT at TREC 2011: Evaluation of Query Expansion Techniques for Medical Record Retrieval. *Proceedings of TREC*, 2011.

[Got12] G. Goth. Analyzing medical data. *Communications of the ACM*, 55(6):13-15, June, 2012.

[Gri64] M. C. Grignetti. A note on the entropy of words in printed English. *Information and Control*, 7:304-306, 1964.

[Hei01] D. T. Heinze, M. L. Morsch, and J. Holbrook. Mining free-text medical records. *Proceedings of AMIA Annual Symposium*, 2001, 254–258.

[Her09] W. Hersh. *Information retrieval: A health and biomedical perspective*, 3rd ed. Springer, 2009.

[Hua12] H.-H. Huang, C.-C. Lee, and H.-H. Chen. Outpatient department recommendation based on medical summaries. *Proceedings of the Eighth Asia Information Retrieval Societies Conference*, LNCS 7675, 518-527.

[Jen06] L. J. Jensen, J. S. and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7:119-129, February 2006.

[Koo11] B. Koopman, M. Lawley and P. Bruza. AEHRC & QUT at TREC 2011 Medical Track: A Concept-Based Information Retrieval. *Proceedings of TREC*, 2011.

[Ram11] P. Ramos. *Acute myocardial infarction patient data to assess healthcare utilization and treatments*. ProQuest, UMI Dissertation Publishing, 2011.

[Sha50] C. E. Shannon. Prediction and entropy of printed English. *Bell System Tech. J*, 30(1):50-64, 1950.

[Voo12] E. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. *Proceedings of TREC*, 2012.

[Voo11] E. Voorhees and R. Tong. Overview of the TREC 2011 Medical Records Track. *Proceedings of TREC*, 2011.