Mining Professional Knowledge from Medical Records

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

Department of Computer Science and Information Engineering, National Taiwan University
#1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
{hhhuang,cclee}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

Abstract. The paper aims at two tasks of electronic medical record (EMR) processing: EMR retrieval and medical term extraction. The linguistic phenomena in EMRs in different departments are analyzed in depth including record size, vocabulary, entropy of medical languages, grammaticality, and so on. We explore various techniques of information retrieval for EMR retrieval, including five retrieval models with six pre-processing strategies on different parts of EMRs. The learning to rank algorithm is also adopted to improve the retrieval performance. Finally, our retrieval model is applied to extract medical terms from EMRs. Both coarse-grained relevance evaluation on department level and fine-grained relevance evaluation on treatment level are conducted.

Keywords: Learning to Rank, Medical Record Retrieval, Professional Information Access.

1 Introduction

Electronic medical records (EMRs) are a special kind of text corpus written by physicians. Medical text mining aims at extracting knowledge from EMRs, constructing a knowledge base (semi-)automatically, and finding new knowledge [1]. Mining medical text from an EMR database is important for case study. The course and treatments of similar cases provide important references, in particular, for medical students or junior physicians. There are many potential applications, e.g., comorbidities and disease correlations [2], acute myocardial infarction mining [3], assessment of healthcare utilization and treatments [4], outpatient department recommendation [5], 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 [5-6]. 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

D. Ślęzak et al. (Eds.): BIH 2014, LNAI 8609, pp. 152–163, 2014.

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.

How to retrieve relevant EMRs effectively and efficiently is an essential research topic. TREC 2011 [7] and 2012 [8] Medical Records track provides test collections for patient retrieval based on a set of clinical criteria. Several approaches such as concept-based [9], query expansion [10], and knowledge-based [11] 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 EMRs, and propose examination, medicine and surgery that may be performed for the input case. Both basic IR models and learning to rank models are explored and discussed.

The structure of this paper is organized as follows. The characteristics of the domain-specific dataset are addressed and analyzed in Section 2. The basic retrieval models and the learning to rank approach are explored in Section 3. Section 4 describes the medical term extraction model and the finer-grained relevance evaluation on course and treatment level. Finally, Section 5 concludes the remarks.

2 An Electronic Medical Record Dataset

The experimental materials come from National Taiwan University Hospital (NTUH). There are 113,625 EMRs in the NTUH dataset. Each EMR is composed of three major parts – say, 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. A course and treatment describes the treatment processes and the treatment outcomes in detail, where medication administration, inspection, and surgery are recorded.

There are 113,625 EMRs 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 EMR 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. In general, the brief history is the longest, while the chief complaint is the shortest.

The 113,625 EMRs 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. 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 3 summarizes the statistics. Compared with the word entropy for general English, the entropy of the medical language used in NTUH dataset is 11.15 bits per word, a little smaller than Shannon entropy (i.e., 11.82 bits per word) [12] and larger than Grignetti entropy (i.e., 9.8 bits per word) [13]. 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. That makes medical record retrieval more challenging.

Table 1. Mean and Standard Deviation of NTUH EMRs 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
EMR	351.62	248.51

Table 2. Distribution of the NTUH EMRs 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. 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 EMR Retrieval

Given a chief complaint and/or a brief history, physicians plan to retrieve the similar cases from the historical EMRs and reference to the possible course and treatments. Chief complaints and/or brief histories in the historical EMRs can be regarded as queries. Section 3.1 describes the basic models and Section 3.2 shows the experimental results. Section 3.3 introduces learning to rank [14] to EMR retrieval. Section 3.4 shows the results and compares them with the basic IR models.

3.1 Basic Models for EMR Retrieval

Words may be stemmed and stop words may be removed before indexing. Spelling checker is introduced to deal with spelling errors and typos. Besides words, medical terms are also recognized as indices. Different IR models can be explored on different parts of EMRs. In the empirical study, Lemur Toolkit is adopted and five retrieval models including TF-IDF, Okapi, KL-divergence, cosine similarity, and indri are experimented.

3.2 Results of the Basic Retrieval Models

In the experiments, 10-fold cross validation is adopted. Given a chief complaint, the output is the retrieved top-n EMRs. We aim to evaluate the quality of the returned n EMRs. 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.

Five retrieval models with six strategies (S1)-(S6) defined as follows are explored.

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

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.

Top 5 and Top 10 EMRs are retrieved and compared. Table 4 shows the experimental results. Overall, the performance tendency is Okapi > TF-IDF > cosine > 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. The Okapi retrieval model with strategy S3 performs the best. 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 other models.

We further evaluate the retrieval models with precision@k shown in Table 5. 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 Fig 1. 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 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 3, 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 4. MAP and nDCG of Basic Retrieval Models 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
cosine	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
cosine	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

Table 5. 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
cosine		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
cosine		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**	**0.6050**	**0.5973**	**0.5965**
KL		0.5775	0.5719	0.5770	0.5725	0.5842	0.5838
cosine		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

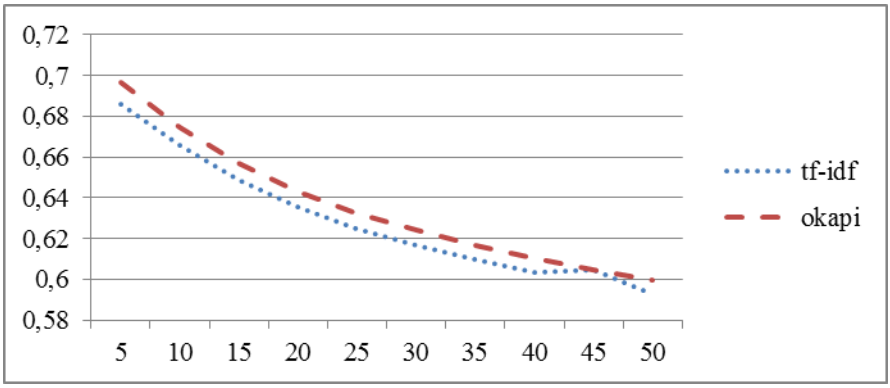

Fig. 1. MAPs of TF-IDF and Okapi under Different *n*'s

Table 6. 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

3.3 Ranking Models for EMR Retrieval

In addition to the fundamental retrieval models, we adopt the learning to ranking model to retrieve the EMRs. Assume a training set is composed of N medical records. Each medical record is regarded as a query. For each query q_i, we retrieval top 200 medical records, m_1, m_2, …, m_{200}, with an IR model. Then, we extract features between q_i and m_1, q_i and m_2, …, q_i and m_{200}. SVM rank along with these features is employed to learn a ranking model. We will use it to re-rank the initial retrieval results.

Table 7. Performance of Learn-to-Rank EMR Models

Model	Metric	BOW	MT	SYMP
TF-IDF	MAP	**0.6989**	**0.6997**	**0.7013**
	nDCG	**0.7620**	**0.7628**	**0.7640**
Okapi	MAP	†0.6957	†0.6964	*0.6992
	nDCG	†0.7593	†0.7597	*0.7618
KL	MAP	0.6970	†0.6968	†0.6977
	nDCG	†0.7595	†0.7592	†0.7604
cosine	MAP	†0.6939	†0.6933	†*0.6968
	nDCG	†0.7571	†0.7565	†*0.7597
indri	MAP	†0.6875	†0.6935	†*0.6954
	nDCG	†0.7516	†0.7567	†*0.7585

In the experiments shown in Section 3.2, the methodology of bag-of-words is adopted. Here we explore two more feature sets: medical terms and symptoms. The medical terms such as examination, medicine, and surgery are extracted from the course and treatment of the retrieved medical records. We describe the details of medical term extraction in Section 4.1.

Physicians often use some fixed patterns to describe symptoms. The following shows some examples for the JJ+NN+NN pattern: left breast pain, congenital heart disease, and bilateral neck mass. We formulate 20 common patterns as follows manually to capture symptoms: (1) JJ NN (and) NN NN, (2) JJ NN NN, (3) JJ (of) NN, (4) VBD NN, (5) NN NN (and) NN, (6) NN (and) NN, (7) JJ VBG NN, (8) VBG NN, (9) JJ NN (and) VBG, (10) NN (of) NN, (11) JJ NN, (12) JJ FW, (13) JJ VBG (and) VBG, (14) VBG (with) NN, (15) NN NN, (16) JJ VBG, (17) JJ JJ NN, (18) NN (with) VBG, (19) NN VBG, (20) NN. The longest-first strategy is adopted.

3.4 Results of the Ranking Models

Table 7 shows the performance of learning-to-rank electronic medical record retrieval models. Top 5 EMRs are retrieved and compared. BOW, MT, and SYMP denote bag-of-words, medical terms, and symptoms, respectively. From column part, using symptoms is better than using bag-of-words and using medical terms, where * denotes SYMP is better than BOW with $p<0.05$. From raw part, TF-IDF model is better than the other four models, where † denotes 95% confidence.

4 Medical Term Extraction

In Section 3, our methods for retrieving EMRs are shown. In addition to the evaluation at department-level, we extract the medical terms such as examination, medicine, and surgery from the course and treatment of the retrieved EMRs. This section shows our extraction models and their performances.

4.1 Extraction Models

To extract the relevant medical terms form EMR, the technology of medical term recognition [15] is required. In this work, 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 [16-17]. 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 the pattern-based approach to extract frequent patterns from medical record dataset and apply them to recognize medical terms. The overall procedure 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.

4.2 Results and Discussion

We evaluate the performance of medical term extraction as follows. 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 8 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

Table 8. Jaccard's Coefficients of Basic and Ranking Retrieval Models on the Course and Treatment Level

Strategy	Top-1	TF-IDF	Okapi	KL	cos	indri
S1	examination	0.3332	0.3448	0.4351	0.3362	0.4501
	medicine	0.2501	0.2995	0.2222	0.2846	0.2035
	surgery	0.1115	0.1406	0.0847	0.1358	0.0776
S2	examination	0.3109	0.3376	0.4017	0.3305	0.4202
	medicine	0.2445	0.2980	0.2370	0.2865	0.2257
	surgery	0.1154	0.1397	0.0961	0.1393	0.0898
S3	examination	0.3515	0.3499	0.4399	0.3437	**0.4535**
	medicine	0.2589	0.3000	0.2245	0.2897	0.2055
	surgery	0.1131	0.1394	0.0844	0.1339	0.0764
S4	examination	0.3289	0.3447	0.4076	0.3362	0.4259
	medicine	0.2539	0.2988	0.2389	0.2905	0.2267
	surgery	0.1168	0.1406	0.0950	0.1376	0.0879
S5	examination	0.3728	**0.3816**	0.3690	0.3814	0.3639
	medicine	0.3166	0.3289	0.3112	0.3292	0.3042
	surgery	0.1851	**0.1954**	0.1821	0.1882	0.1758
S6	examination	0.3727	0.3810	0.3679	0.3826	0.3636
	medicine	0.3147	0.3278	0.3101	0.3291	0.3035
	surgery	0.1835	0.1936	0.1803	0.1875	0.1743
Ranking	examination	0.3852	0.3852	0.3847	0.3890	0.3846
	medicine	0.3291	0.3301	0.3310	**0.3348**	0.3313
	surgery	0.2012	0.2007	**0.2013**	0.2005	0.1999

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 seven strategies used in the above experiments are explored again in the fine-grained evaluation. S1 to S6 are based on the basic retrieval models describe in Section 3.2. Ranking is the learning-to-rank model with symptom features shown in Section 3.3. Overall, the performance of examination prediction is larger than that of medicine prediction, which is larger than that of surgery prediction in all models. Considering brief history (i.e., strategies S5 and S6) benefits medicine and surgery prediction. Excluding the learning to rank approach, the 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). In other words, the information from brief history induces noises in examination prediction. The Comparison between S5 and the Ranking shows that the learning to rank approach improves the performances on all the examination, medicine, and surgery predictions in all the five models.

5 Conclusion

This paper aims at mining the professional knowledge from medical records. We compare different retrieval models under different strategies on department and course and treatment levels. In addition, ontology-based and pattern-based approaches are adopted to extract medical terms. Both coarse-grained and fine-grained relevance evaluations with various metrics are conducted.

Some linguistic phenomena in EMRs are identified. 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.

In the experiments of basic retrieval models, five retrieval models and six index strategies are tested. The Okapi model achieves the best performance. Query accesses to the medical records in medical languages of smaller entropy tend to have better performance. The performance of departments related to generic parts of body such as Department of Oncology and Department of Neurology are worse than average performance. In the experiments of the learning to rank approach, we explore the ranking approach on five retrieval models and three index strategies. Under the learning to rank algorithm, the TF-IDF model with the symptoms strategy achieves the best performance. Applying learning to rank technique is significantly better than those models.

Our mining approach can be adopted in various applications. The medical record retrieval can be applied to create a search engine that delivers similar medical records for education and case study. The outpatient recommendation system is another application. For example, a patient can search for the appropriate outpatient department by

inputting the patient's chief complaints. The medical term extraction can be applied to analyze the correlations between drug and symptoms. Moreover, a medical assistance system can be constructed to detect the anomalous treatments and remind the physicians to double-check their diagnosis.

Acknowledgements. This research was partially supported by Ministry of Science and Technology, Taiwan, under the grant 101-2221-E-002-195-MY3.

References

1. Jensen, L.J., Saric, J., Bork, P.: Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics 7, 119–129 (2006)
2. Goth, G.: Analyzing medical data. Communications of the ACM 55(6), 13–15 (2012)
3. Heinze, D.T., Morsch, M.L., Holbrook, J.: Mining free-text medical records. In: AMIA Annual Symposium, pp. 254–258 (2001)
4. Ramos, P.: Acute myocardial infarction patient data to assess healthcare utilization and treatments. ProQuest, UMI Dissertation Publishing (2011)
5. Huang, H.-H., Lee, C.-C., Chen, H.-H.: Outpatient department recommendation based on medical summaries. In: Hou, Y., Nie, J.-Y., Sun, L., Wang, B., Zhang, P. (eds.) AIRS 2012. LNCS, vol. 7675, pp. 518–527. Springer, Heidelberg (2012)
6. Hersh, W.: Information retrieval: A health and biomedical perspective, 3rd edn. Springer (2009)
7. Voorhees, E., Tong, R.: Overview of the TREC 2011 Medical Records Track. In: TREC (2011)
8. Voorhees, E., Hersh, W.: Overview of the TREC 2012 Medical Records Track. In: TREC (2012)
9. Koopman, B., Lawley, M., Bruza, P.: AEHRC & QUT at TREC 2011 Medical Track: A Concept-Based Information Retrieval. In: TREC (2011)
10. Dinh, D., Tamine, L.: IRIT at TREC 2011: Evaluation of Query Expansion Techniques for Medical Record Retrieval. In: TREC (2011)
11. Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Rance, B., Lang, F., Ide, N., Apostolova, E., Aronson, A.R.: A Knowledge-Based Approach to Medical Records Retrieval. In: TREC (2011)
12. Shannon, C.E.: Prediction and entropy of printed English. Bell System Tech. J. 30(1), 50–64 (1950)
13. Grignetti, M.C.: A note on the entropy of words in printed English. Information and Control 7, 304–306 (1964)
14. Li, H.: A Short Introduction to Learning to Rank. IEICE Trans. Inf. & Syst. E-94D(10), 1–9 (2011)
15. Abacha, A.B., Zweigenbaum, P.: Medical entity recognition: a comparison of semantic and statistical methods. In: Workshop on Biomedical Natural Language Processing, pp. 56–64 (2011)
16. Chen, H.-B., Huang, H.-H., Chen, H.-H., Tan, C.-T.: A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications. In: 24th International Conference on Computational Linguistics, pp. 545–560 (2012)
17. Chen, H.-B., Huang, H.-H., Tjiu, J., Tan, C.-T., Chen, H.-H.: A statistical medical summary translation system. In: ACM SIGHIT International Health Informatics Symposium, pp. 101–110 (2012)