MKDS: A Medical Knowledge Discovery System Learned from Electronic Medical Records (Demonstration)

Hen-Hsen Huang[1(✉)], An-Zi Yen[1], and Hsin-Hsi Chen[1,2]

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

[2] MOST Joint Research Center for AI Technology and All Vista Healthcare, National Taiwan University, Taipei, Taiwan

Abstract. This paper presents a medical knowledge discovery system (MKDS) that learns the medical knowledge from electronic medical records (EMRs). The distributed word representations model the relations among medical concepts such as diseases and medicines. Four tasks, including spell checking, clinical trait extraction, analogical reasoning, and computer-aided diagnosis, are demonstrated in our system.

Keywords: Medical knowledge discovery · Medical records
Distributed word representation

1 Introduction

Clinical decision supporting (CDS) systems provide physicians with professional knowledge for clinical decision-making [3]. The popular CDS systems such as UpToDate and Micromedex are online database systems containing the information of drugs, diseases, diagnosis, symptoms, exams, surgeries, and so on. After a user enters a keyword, e.g., "leukemia", the prognosis, the symptoms, the exams, and the treatments for each subtype of leukemia are shown. Such information is very useful for physicians and students to do clinical case studies.

On major CDS systems, the contents are human edited. Domain experts organize medical information into a database based on their prior knowledge. In this paper, we show an approach to discover the relations among medical concepts from medical documents, and apply the medical knowledge to aid some medical applications.

Electronic medical records (EMRs) written by physicians is a rich source of professional knowledge [3]. This work presents a medical knowledge discovery system (MKDS)[1], which aids to discover the relationships among medical concepts from EMRs. Different from the approaches based on traditional information retrieval [9], our method utilizes the skip-gram model, which has been shown to represent lexical term

[1] http://nlg18.csie.ntu.edu.tw:8181.

Y.-H. Tseng et al. (Eds.): AIRS 2018, LNCS 11292, pp. 196–202, 2018.
https://doi.org/10.1007/978-3-030-03520-4_19

semantics effectively [8], to capture the relatedness among medical concepts. We demonstrate the uses of word vectors in four tasks in our system, including spell checking, clinical trait extraction, analogical reasoning, and computer-aided diagnosis. Compared to traditional approaches, our method addresses healthcare issues like literature analysis, prognosis, and patent management.

The main contributions of this work are three folds: (1) We show how the word vectors learned from EMRs can help capture medical relationships among medicines, surgeries, diagnosis, and exams. (2) Four applications based on the extracted knowledge are demonstrated with an instructive and educational system. (3) We release word vectors trained from medical documents[2]. That can be applied to various clinical NLP applications.

2 Related Work

Knowledge discovery in medical documents is an attractive topic. The medical tracks in TREC 2011 and 2012 deal with the task of information retrieval on EMRs [16]. In i2b2 NLP challenges, the shared tasks, including medical term extraction and classification of relations between medical concepts, are explored with EMRs [12]. EMRs are used in many applications including assertion classification [6], clinical trait extraction [2], co-reference resolution [14], temporal analysis [13], pneumonia identification [1], phenotyping [4, 10, 11], and outpatient department recommendation [5].

Neural network models such as word2vec [8] represent a word as a vector in a low-dimensional space, where semantic relatedness can be measured. In addition, the property of linguistic regularity is also observed [7, 9]. Using distributed word representation is also shown to improve the clinical concept extraction [15].

3 Resources

The experimental dataset is composed of the EMRs from National Taiwan University Hospital (NTUH). As shown in Table 1, total 113,625 medical records are collected in five years. An EMR contains three main sections, the chief complaint, the brief history, and the course and treatment. The chief complaint denotes the purpose of the patient's

Table 1. Statistics of the EMRs in the NTUH corpus.

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

[2] http://nlg18.csie.ntu.edu.tw/mkds/medical.w2v.

visit, e.g., “headache for a week”. The brief history presents the background information of the patient like her/his age, gender, and past diseases. The course and treatment note the treatment such as medicines, surgeries, and exams for the patient.

The Unified Medical Language System (UMLS) is adopted as the ontology of medical terms. Five types of medical terms including drug, exam, surgery, diagnosis, and body are collected from the UMLS.

4 Medical Term Representation

We learn the distributed word representations from the preprocessed EMRs.

4.1 Preprocessing

EMRs are usually written in a rush, thus noise is unavoidable. Spelling errors, grammatical errors, and abbreviations are common in EMRs. Preprocessing is performed to deal with the related issues.

Age. Six age patterns are found in EMRs: “#-year-old”, “#-years-old”, “# year old”, “# years old”, “# y/o”, and “#-y-o”. All the ages are standardized to four groups: under 15 years old, 16 to 45 years old, 46–60 years old, and beyond 60 years old.

Gender. Eight gender patterns are found in the NTUH corpus: woman, lady, female, girl, man, gentleman, male, and boy. We standardize them to two types of gender.

Medical Terms. In EMRs, the UMLS terms are identified as medical terms. The phrases and the compounds such as “coronary artery disease” are treated as single terms. For each of major diseases such as “diabetes” or “coronary artery disease”, we merge all its alias and abbreviations into one. For each drug, we replace all its trade names with the generic name, and discard the dosage form.

4.2 Representation Learning

The Skip-gram model is used to train the word representations of medical terms. The size of dimension is 500, the context window is 10, and negative sampling is used with size 5. In this way, a medical term t is represented by a vector v with length 500, and the relatedness between two medical terms t_1 and t_2 can be measured by the cosine similarity of their vectors v_1 and v_2.

5 Medical Knowledge Discovery

The learned medical word vectors are employed to four tasks in our system.

5.1 Spell Checking

We collect the frequent OOV terms from the NTUH corpus. We check if an OOV term t has spelling error, and correct it by using the medical word vectors. First, we fetch 10

medical terms most related to *t*. The longest common subsequence (LCS) algorithm is performed to measure the overlap between a candidate *c* and *t*. The most related candidate *c* with an overlap of at least 90% is chosen as the counterpart of *t*. Table 2 lists the top 5 spelling errors. The homophone error "leukovorin" is corrected. In EMRs, the exam colonscopy is often misspelled as colonscope, which is the instrument used in the exam. Full list is available on the website of MKDS.

Table 2. Most frequent spelling errors in the NTUH corpus.

Error	Correction	# Occurrences
Leukovorin	Leucovorin	6,013
Lower extremity	Lower extremities	2,308
Colonscope	Colonscopy	1,801
Esophageal varix	Esophageal varices	1,709
Hypermetabolism	Hypermetabolic	1,391

5.2 Clinical Trait Extraction

MKDS extracts the most related medical terms in each type for an input term. To evaluate the performance, domain knowledge is consulted by the national cancer institute website (NCI). Six common cancer types, including lung cancer, female breast cancer, colon cancer, prostate cancer, esophageal cancer, and leukemia, are chosen. From the NCI data, the drugs approved to treat each type of cancers are listed. These medications form the ground truth for evaluation. Because not all the drugs listed on NCI are used in NTUH, the intersections of NTUH and NCI are computed.

The performance is measured by average precision at 10 (ap@10). Table 3 shows the drugs suggested by our system for each type of cancers. Different from the NCI list, our clinical traits are weighted. Among 63 drugs for breast cancer, the top one, tamoxifen citrate, is the most common hormone therapy of the estrogen receptor positive (ER+) breast cancer. For lung cancer, gefitinib is a kind of targeted therapy for non-small cell lung cancers with mutated epidermal growth factor receptor (EGFR). Gefitinib is a common drug for lung cancer treatment in NTUH because EGFR mutations are much more prevalent in Asia.

Table 3. Performance of clinical trait extraction.

Cancer	ap@10	Approved drugs for treatment
Lung	64.15%	Gefitinib, erlotinib, pemetrexed disodium heptahydrate
Breast	42.19%	Tamoxifen citrate, zoladex, halcion, simethicone, goserelin
Colon	53.14%	Oxaliplatin, cetuximab, capecitabine, bevacizumab
Prostate	43.19%	Flutamide, bicalutamide, deflux
Esophageal	12.86%	Docetaxel
Leukemia	39.88%	Imatinib mesylate, idarubicin, daunorubicin, cytarabine

5.3 Analogical Reasoning

Analogical reasoning based on word vectors can be simplistically accomplished by vector arithmetic like 3COSADD [9]. Our system answers the analogy questions like t_1:$t_2 \approx t_3$:t_4, where t_4 is unknown. For example, "tamoxifen citrate" is one of answers to the question "lung cancer:gefitinib $\approx$ breast cancer:?".

Referred to Table 3, we compose questions t_1:$t_2 \approx t_3$ with the top one medication for each of the six cancer types. Table 4 shows some promising analogies. The relation between a cancer type and its drug can be captured by vector offsets.

Table 4. Analogical reasoning between cancers and medications.

t_1 (cancer)	t_2 (drug)	t_3 (cancer)	t_4 (drug)
Prostate	Flutamide	Lung	Gefitinib
Prostate	Flutamide	Breast	Tamoxifen citrate
Leukemia	Imatinib mesylate	Colon	Oxaliplatin
Lung	Gefitinib	Leukemia	Imatinib mesylate
Breast	Tamoxifen citrate	Lung	Gefitinib

5.4 Computer-Aided Diagnosis

In this task, a user inputs a chief complaint, and the system responses the most related diagnosis and treatment. Rather than a single medical term, a chief complaint is a short sentence or a statement, where medical and non-medical terms are intermixed. Figure 1 shows a snapshot of the computer-aided diagnosis in MKDS.

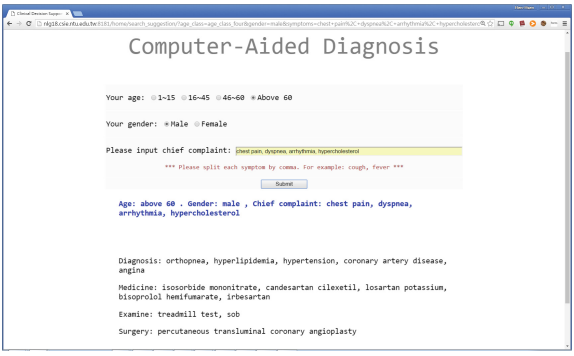

Fig. 1. Snapshot of the use of computer-aided diagnosis.

6 Conclusions

This work demonstrates four applications of distributed medical term representations. The properties of the skip-gram model such as semantic relatedness and linguistic regularity are utilized to implement the four tasks. Compared to the traditional pre-trained word vectors, our approach acquires the knowledge from the real-world medical records. With more and newer EMRs, our system will be more instructive.

Acknowledgements. This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 106-3114-E-009-008 and MOST-105-2221-E-002-154-MY3, and National Taiwan University under grant NTUCCP-106R891305.

References

1. Bejan, C.A., Vanderwende, L., Wurfel, M.M., Yetisgen-Yildiz, M.: Assessing pneumonia identification from time-ordered narrative reports. In: Proceedings of 2012 AMIA Annual Symposium, pp. 1119–1128 (2012)
2. Davis, M.F., Sriram, S., Bush, W.S., Denny, J.C., Haines, J.L.: Automated extraction of clinical traits of multiple sclerosis in electronic medical records. J. Am. Med. Inform. Assoc. **20**(2), 334–340 (2013)
3. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.: What can natural language processing do for clinical decision support? J. Biomed. Inform. **42**(5), 760–772 (2009)
4. Hripcsak, G., Albers, D.J.: Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. **20**(1), 117–121 (2013)
5. Huang, H.-H., Lee, C.-C., Chen, H.-H.: Mining professional knowledge from medical records. In: Ślęzak, D., Tan, A.-H., Peters, James F., Schwabe, L. (eds.) BIH 2014. LNCS (LNAI), vol. 8609, pp. 152–163. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09891-3_15
6. Kim, Y., Riloff, E., Meystre, S.M.: Improving classification of medical assertions in clinical notes. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers, pp. 311–316 (2011)
7. Levy, O., Goldberg, Y.: Linguistic regularities in sparse and explicit word representations. In: Proceedings of the 18th Conference on Computational Language Learning, pp. 171–180 (2014)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop Papers (2013)
9. Mikolov, T., Yih, W.T., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT, pp. 746–751 (2013)
10. Pathak, J., Kho, A.N., Denny, J.C.: Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J. Am. Med. Inform. Assoc. **20**(e2), e206–e211 (2013)
11. Shivade, C., et al.: A review of approaches to identifying patient phenotype cohorts using electronic health records. J. Am. Med. Inform. Assoc. **21**, 221–230 (2014)
12. Stubbs, A., Kotfila, C., Xu, H., Uzuner, Ö.: Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task track 2. J. Biomed. Inform. **58**, S67–S77 (2015)

13. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J. Am. Med. Inform. Assoc. **20**(5), 806–813 (2013)
14. Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., South, B.R.: Evaluating the state of the art in coreference resolution for electronic medical records. J. Am. Med. Inform. Assoc. **19**(5), 786–791 (2012)
15. De Vine, L., Kholghi, M., Zuccon, G., Sitbon, L., Nguyen, A.: Analysis of word embeddings and sequence features for clinical information extraction. In: Proceedings of the 13th Annual Workshop of the Australasian Language Technology Association (2015)
16. Voorhees, E.M., Hersh, W.: Overview of the TREC 2012 medical records track. In: Proceedings of the 21st Text REtrieval Conference (2012)