Keyword Extraction and Structuralization of Medical Reports

Ching-Wei Tsai
Chinese Medical University Hospital and College of Medicine
Pei-Hao Wu
National Taiwan Normal University
Avon Yu
National Taiwan Normal University
Jia-Ling Koh
National Taiwan Normal University
Chin-Chi Kuo
China Medical University Hospital and College of Medicine
Arbee L.P. Chen (✉ **arbee@asia.edu.tw**)
Asia University https://orcid.org/0000-0003-2872-4484

Research article

Keywords: medical reports, keyword extraction, medical text reports

Posted Date: August 4th, 2019

DOI: https://doi.org/10.21203/rs.2.11643/v1

Abstract

In recent years, patients usually accept more accurate and detailed examinations because of the rapid advances in medical technology. Many of the examination reports are not represented in numerical data, but text documents written by the medical examiners based on the observations from the instruments and biochemical tests. If the above-mentioned unstructured data can be organized as a report in a structured form, it will help doctors to understand a patient's status of the various examinations more efficiently. Besides, further association analysis on the structuralized data can be performed to identify potential factors that affect a disease. In this paper, from the pathology examination reports of renal diseases, we applied the POS tagging results of natural language analysis to automatically extract the keyword phrases. Then a medical dictionary for various examination items in an examination report is established, which is used as the basic information for retrieving the terms to construct a structured form of the report. Moreover, a topical probability modeling method is applied to automatically discover the candidate keyword phrases of the examination items from the reports. Finally, a system is implemented to generate the structured form for the various examination items in a report according to the constructed medical dictionary. The results of the experiments showed that the methods proposed in this paper can effectively construct a structural form of examination reports. Furthermore, the keywords of the popular examination items can be extracted correctly. The above techniques will help automatic processing and analysis of medical text reports.

Background

With the advance of technology, medical and healthcare data continues to rise which is easier to collect. How to discover useful information from the targeted medical data in order to further optimize various medical services has become one of the important research topics. The electronic medical records (EMR) system provides an environment for storing all of the health information of the patients in an electronic format. The data includes the descriptions on medication, laboratory test results, and symptoms of patients recorded by the doctors. Accordingly, the analysis on this type of records will be a direct way to understand the progress of a disease, which can assist healthcare professionals to make accurate and efficient medical treatment for patients.

Recent advancement in medical technology allows for more meticulous and detailed examinations for patients. However, many of the examination reports are not represented in numerical form. Instead, the reports contain text descriptions for the results shown in a medical equipment and observations of technical inspection provided by a medical technologist. From these unstructured examination reports, it often requires time for a doctor to read and understand what unusual problems occur in order to diagnose medical conditions and propose the proper treatments. Converting an unstructured text examination report into a structured one can help doctors understand a patient’s conditions for various examined items more effectively and efficiently. Moreover, it facilitates the doctor's breakdown of the different classification criteria for medical records and further analyzes the relevance of clinical symptoms to identify potential factors affecting the disease. Accordingly, the technology of transferring an unstructured examination report into a structured one is important and in practical needs, which can further improve the working efficiency and quality for the healthcare professionals.

The research goal of this paper is to construct a system for converting an unstructured English medical report into a structured form to display. In this study, we studied the nephrology medical examination reports. Sethi and Sanjeev et. al [17] proposed that a nephrology medical examination report should include the following eight paragraphs, which contain the different examination items and contents, respectively: 1) Main diagnosis (Diagnosis), 2) Electron Microscopy Examination (EM), 3) Examination status of electron microscope (Comment/Narrative), 4) The size and numbers of the sliced specimen (Specimen type), 5) General description of the spliced specimen (Gross description), 6) Examination of Light Microscope (LM), 7) Chromosome examination (DIF), and 8) Report Conclusion (Summary). Within an examination report, if the contents correspond to the above paragraphs, which are surely follow the certain order. Paragraph (1), (3) and (8) are abstracts for the symptoms and examination results summarized by a doctor. The goal of processing these paragraphs is to extract the main diagnosis including the judgment of the main disease and the descriptions for abnormality of the pathological examines. Moreover, paragraph (2), (4), (5), (6), and (7) are special inspection contents for the various inspection items observed and recorded. The goal of processing is to indicate the resultant options for a specific inspection items. Figure 1 shows an example of a nephrology examination report, the content in the report is separated according to the paragraphs. The strategies proposed in this paper will automatically extract the result of each examination item as shown in Table 1 (in the Supplementary Files).

The data set used in this study consists of 476 examination reports of anonymous patients in the Department of Nephrology of China Medical University Hospital. All the contents of the report are written in English. The contribution of this paper are as follows:

We design a keyword phrase extraction method, which can be used to construct a vocabulary dictionary from the clinical examination reports automatically.

A strategy is proposed to automatically discover the keyword terms which represent the examination items from the examination reports.

According to the constructed vocabulary dictionary, a system prototype is constructed to organize the results of the various examination items in structured form.

The processing of the proposed system includes two parts: the offline training and the online processing. The offline processing is shown in Figure 2, in which the reports are inputted to a natural language processing tool in batches to get their part-of-speech tagging at first. Then a medical dictionary is built for each paragraph in the report. Besides, another module is proposed for automatically discovering the keyword terms of examinations terms by analyzing the topical terms in each paragraph of an examination report. The online processing is shown in Figure 3, in which an examination report will be converted into a structured form of report based on the dictionary and keyword phrases extracted from the offline training results.

At first, each examination report is segmented into paragraphs. According to the established vocabulary dictionary, the continuous vocabulary words are combined into keyword phrases according to their part-of-speech tags in the content of paragraphs. Then the words with typos can be corrected by their similar keywords with higher frequencies. To perform the key terms extraction, each paragraph in a report is assumed to describe a specific topic content for the examination. Therefore, the LDA statistical generative model is used to discover the representative topic words for each paragraph as candidate keywords of the examination items. By computing the entropy of the topic words in the various paragraphs of the report, the general terms and specific terms of the examination items are distinguishable. The specific item terms are extracted accordingly, which are used as binary attributes to represent the examination result with a structured form.

The following sections of this paper are organized as follows: Section 2 describes the related works. Section 3 explains the method for constructing the medical vocabulary dictionary and introduces the strategies of discovering the keyword terms of examination items, which are used to show the examination results in a structured form. Section 4 provides the evaluation results of the proposed system. Finally, Section 5 concludes this paper and discusses the future work.

Related Works

With the advance of information technology, medical records can be collected and maintained easily. Therefore, the clinical data of a lot of patients are stored in a database in electronic form such as medical diagnoses, medications, and examination reports. How to apply the data mining strategies properly to find association information from medical records, which can be used as a decision-making reference, has become an attracting research direction in recent years.

Unstructured clinical notes contain a wealth of information about each patient, but extracting it is a difficult problem by its complete lack of structure. This makes deriving information about patient characteristics from clinical notes a computationally challenging task that requires sophisticated NLP (Natural Language Processing) tools and techniques such as cTAKES [16] and MetaMap [1].

Information extraction has been exploited in some clinical research domains [20], studies show mostly used clinical IE approaches are rule-based and machine learning-based. One of a common form of rule-based method is based on regular expressions, which many of the defined search patterns are recognized and written manually. Savova et al. [15] used regular expressions to identify peripheral arterial disease (PAD). A positive PAD was extracted when the predefined patterns were matched.

Machine learning-based methods have gained attention because of their effectiveness. Nandhakumar et al. [12] use a word-level, sentence-level features and applied Conditional Random Fields (CRFs) model [8] to extract the clinically significant parts of the radiology reports. Then, the reports are classified into critical or non-critical categories which help physicians to identify high priority reports that need urgent treatment. [7] proposed a mortality prediction for the patients in the intensive care units in order to make a most appropriate decision. The authors believe that the nursing notes within a recent time period can identify hidden clues about the physical condition of a patient, which is useful to decide the priority of handling matters. A state transition topic model is established to capture the semantic information in a nursing note. Then the n-grams, standard topics, state-aware topics, states etc. are used as the extracted features for a cost-sensitive SVM (Support Vector Machine) classifier to perform mortality prediction. [5, 10] also considered the problem of mortality prediction from the nursing notes. The former [5] used LDA (Latent Dirichlet allocation) [3] to decompose free-text notes into meaningful features, after that, the SVM classifier is used to predict mortality. On the other hand, the latter [10] used topic model distribution as features, then a logical linear regression method is used to predicting mortality probability.

[6] used electronic medical records (EMRs) to automatically construct a medical knowledge graph in hope to help improving the treatment decision-making for patients. For a given medical question, the answer is first discovered and then the scientific articles containing the answer is selected and ranked. The proposed method establishes the information in an electronic medical record with the Markov network model to create a medical knowledge map. The probability of each connected edge in the pattern structure is computed. Finally, the probability values of possible answers is inferred from a probabilistic knowledge graph, which is automatically generated by processing the unstructured text in a large collection of electronic medical records (EMRs).

[9] believed that medication feedback can be extracted by the posts of patients in a social media site, which is useful to obtain the possible reaction of drugs from the response of patients. The researches of [4, 11] aimed to find out the status and factors of adverse drug reactions from the discussions of the Internet users about the reaction of drug usage. The method proposed in [4] finds out the structure grammar between drugs, symptoms, and diseases in sentences from the discussion of drugs on Internet forum. Then unsupervised relation extraction method is used to discover the domain-specific relation patterns. The post-processing algorithm then merges the missing or incomplete sentences into complete sentences. Finally, from the extracted sentences, the lift measure is applied to evaluate the correlation between a drug and a symptom.

[2] thinks it is a complicated problem to answer what diseases a patient has only according to the textual description of symptoms of the patient. The reason is that the same symptoms may occur in many diseases, which makes it difficult to decide the disease of a patient only from part of the symptoms. Moreover, this property makes it not easy to search the similar cases of a patient. The problems described above show the demand of the research studied in this paper. To get a structured form of a textual examination report, a database of diseases and the clinical features can be established according to the extracted structured content. Accordingly, the doctors can query the diseases by querying different clinical features as conditions, and further analyze the correlation of diseases and results of examination items.

Keyword extraction is a critical technology to get semantic information from examination reports, which aims to obtain the keywords which represent the main points in the report, which are used for subsequent analysis. [14] considered the problem of entity set expansion. The purpose is to find a set of entities by giving one or a few seed examples, which belong to the same semantic class. For example, the entity "apple" is a kind of fruit. Therefore, entity set expansion would like to find the entities such as "banana", "orange" and etc. which are also fruits. To extract the sibling relations from text, one important feature is the Skip-gram because it provides positional constraints on the contextual words with regard to the target term. Accordingly, for a given keyword, the surrounding text is more likely to appear as an extension of the keyword. [19] aimed to extract the general knowledge from the narratives of movies. For example, for a sentence “The man began to shoot a video in the moving bus”, the goal is to discover the facts (“the man”, “began to shoot”, “a video”) and (“the man”, “began to shoot”, “in the moving bus”) described in the sentence. One of the pre-processing steps is to perform text segmentation to extract the syntactic phrases in the sentence. Accordingly, the OpenNLP is used to decompose sentence into (“the man”), (“began to shoot”), (“a video”), (“in”), and (“the moving bus”). Upon completing the above pre-processing, the semantic dependencies between these phrases are analyzed to generate the fact triples. This paper provides some ideas of text segmentation for semantics analysis.

Methods

3.1 of Medical Dictionary Construction

In this section, we will describe the pre-processing on the examination reports in order to establish a dictionary of the medical vocabulary for the reports.

3.1.1 Pre-processing of Examination Reports

In order to establish a dictionary of medical vocabulary, the whole corpus of the examination reports is pre-processed, which includes paragraph segmentation and part-of-speech tagging.

<1> Paragraph Segmentation

As shown in the introduction, a nephrology medical examination report itself has the structural aspects of the 8 paragraphs: 1) Diagnosis, 2) EM, 3) Comment/Narrative, 4) Specimen type, 5) Gross description, 6) LM, 7) DIF, and 8) Summary. The text descriptions of each paragraph are distinct, and the frequencies of the words within them are also different. Therefore, this paper will establish the corresponding dictionaries of different paragraphs. In order to achieve this goal, the contents of the examination report must be automatically segmented according to the patterns appearing at the beginning of the paragraph, as shown in Table 2 (Supplementary Files). After a report is segmented into paragraphs as the result shown in Table 3 (Supplementary Files), which will be proceed to perform part-of-speech tagging.

<2> Part-of-speech Tagging (POS Tagging)

We used the Stanford CoreNLP API, a natural language processing tool (https://stanfordnlp.github.io/CoreNLP) developed by Stanford University's Natural Language Processing Research Group [18], to perform POS tagging. Figure 4 shows the result of part-of-speech tagging on a sentence in a Diagnosis paragraph, which includes the tense of each word in the sentence such as NN (noun), VBD (past tense verb), JJ (adjective), and CD (quantifier). According to the part-of-speech tags of words, the NNS (plural noun), VBG (gerund), NNP (proper noun), etc. are classified as nouns. On the other hand, VBN (past participle), VBD (past tense verb), JJR (comparative adjective), etc. are classified as adjectives.

3.1.2 Medical Dictionary Construction

After the corpus of examination reports are pre-processed, the next step is to construct the medical dictionary for the different paragraphs in a clinical examination report. The detailed processing of constructing the dictionary include the following three steps.

<Step 1>

At first, the pre-processing result of the examination reports are collected, as shown in Table 4 (in the Supplementary Files). Next, the continuous words appearing in the sentences are combined into a vocabulary term according to some specific pattern rules as shown in Table 5 (in the Supplementary Files). Then the extracted vocabulary terms are organized into the following two dictionaries:

Adjective vocabulary dictionary: which includes the adjectives and the compound adjectives (continuous adjectives).

Proper noun vocabulary dictionary: which includes the nouns and the compound nouns. The compound nouns are further divided into: (1) a combination of nouns which uses the last noun word as the base word; (2) the compound noun has an adjective as its prefix word.

<Step 2>

In this step, the extracted phrases in step 1 are refined to filter out the semantics meaningless words according to the following rules: (1) the noun-phrases with medical meaningless ending words are removed, and (2) the adjectives that are medical meaningless at the beginning of the phrases are removed. Some examples of the medical meaningless ending words and adjectives are shown in Table 6 (in the Supplementary Files). These meaningless words are given manually.

<Step 3>

The typos appearing in the clinical reports usually occur due to additional characters or missing characters. In order to filter out the typos in the dictionary, we apply the Longest Common Subsequence [13] algorithm to estimate the similarity between the extracted phrases. If two words are similar to each other, the word with a less frequency is replaced by the other.

Step 3-1: Filter out the typos based on ending words.

Let $b_i$and b_j denote two different base words, which are the last words of the extracted phrases. Because the additional characters or missing characters of typos usually occur in the middle or the ending of a word, the initial characters of b_i and b_j should be the same. Let $max\_len(b_i, b_j)$ denotethe maximal length between b_i and b_j, and $LCS(b_i, b_j)$ denote the length of the longest common subsequence of b_i and b_j. Then the typing error between b_i and b_j is computed by $max\_len(b_i, b_j)$ - $LCS(b_i, b_j)$ and denoted as $ErrBaseW(b_i, b_j)$, as shown in equation 1.

$ErrBaseW(b_i, b_j) = max\_len(b_i, b_j) – LCS(b_i, b_j)$ (Eq. 1)

Next, max_len(b_i, b_j) is multiplied by $1/d$ to get a threshold value, denoted as $ComBaseT(b_i, b_j)$, of the typing error between b_i and b_j, as shown in equation 2.

$ComBaseT(b_i, b_j) = max\_len(b_i, b_j) / d$□ (Eq. 2)

When the typing error is less than or equal to the threshold value $ComBaseT(b_i, b_j)$, b_i and b_j are considered to occur a typing error and $ComBaseF(b_i, b_j)$ is set to be 1; otherwise $ComBaseF(b_i, b_j)$ is set to be 0 as shown in equation 3.

(Eq. 3)

Let B_i and B_j denote the set of phrases whose base words are b_i and b_j, respectively. Besides, $F(b_i)$ and $F(b_j)$ denote the frequencies of b_i and b_j, respectively. If $ComBaseF(b_i, b_j)$ is 1 and $Fb_i > Fb_j$, all the base words in B_j are modified into b_i and are merged with the phrases in B_i.

□Example 3-1□

Assume there are two sets of phrases with different base words, as shown in Table 7 (in the Supplementary Files). At first, check the base words with initial letter '*g'* in pairs. Next, Eq. 1 is used to compute the typing error between the base words: 'glomerulonephritis' and 'glomerulonephritiss'. Because *ErrBaseW* ('glomeru-lonephritis', 'glomerulonephritiss') is 1 and the threshold of the merging the base words *ComBaseT*('glomerulonephritis', 'glomerulonephritiss') is 4 when *d* is set to 5. According to Eq. 3, the typing error 1 is acceptable, it results in the two base words are corrected into the one with a higher frequency. Because F('glomerulonephritis') > F('glomerulonephritiss'), the words in the set with base word 'glomerulonephritiss' is modified to have a base word 'glomerulonephritis'.

Step 3-2: Perform a filtering operation on a set B_i consisting of the phrases with the same base word.

Let P_i and P_j denoted two phrases in the set B_i, which have the same base words. Let $P_i.w_1$,… , $P_i.w_n$ denote the sequence of words in P_i and $P_j.w_1$,… , $P_j.w_m$ denote sequence of words in P_j. If both P_i and P_j

consist of the same number of words, the typing errors between these two phrases are counted. Otherwise, P_i and P_j are considered to be different phrases, as shown in Eq. 4.

$CountContentW(P_i, P_j)$ =1 if n=m, =0 otherwise (Eq. 4)

Eq. 5 is used to compute typing errors between $P_i.w_k$ and $P_j.w_k$ for k = 1 to n. Eq. 6 is used to compute the threshold value of typing errors for w_1 to w_n.

$ErrContentW(P_i.w_k, P_j.w_k) = max\_len(P_i.w_k, P_j.w_k) - LCS(P_i.w_k, P_j.w_k)$ (Eq. 5)

$CountContentT(P_i.w_k, P_j.w_k) = min(max\_len(P_i.w_k, P_j.w_k)\ /\ d$□, 3) (Eq. 6)

If the typing error of any word w_k in P_i and P_j are larger than the threshold $CountContentT(P_i.w_k, P_j.w_k)$, P_i and P_j are considered to be two different phrases, as shown in Eq 7. Otherwise, if the frequency of P_i is larger than P_j, i.e. $F(P_i)$> $F(P_j)$, P_j is considered as a typo of P_i and thus modified to be P_i.

00

(Eq. 7)

□Example 3-2□

Suppose that a set of phrases with the same base word is shown in Table 8 (in the Supplementary Files). Let P_i and P_j denote "mildd tubular atrophy" and "mild tubular atrophy" in the set, respectively. At first, Eq. 4 is used to determine whether “mild tubular atrophy” and “mildd tubular atrophyis” consist of the same number of words. Then for each word in the prefix of the two phrases, the typing error is compared with the error threshold, where $ErrContentW$(‘tubular’, ‘tubular’)=0 and $ErrContentW$(‘mild’, ‘mildd’) =1, and the threshold values $ComContentT$(‘tubular’, ‘tubular’) = 2 and $ComContentT$(‘mild’, ‘mild’) = 1. Because each word in the two phrases satisfying the checking process of typing error, these two phrases are combined together. Because F("mild tubular atrophy") > F("mildd tubular atrophy"), the phrase "mildd tubular atrophy" is modified to be "mild tubular atrophy". Accordingly, the frequency count of "mild tubular atrophy" is updated to be 45.

3.2 Usage of Medical Dictionary

This section introduces how to use the medical vocabulary dictionary to perform structuralization for the medical reports. The following two subsections will introduce the overall processing of structuralization and how to automatically extract the keywords for each special inspection item.

3.2.1 Methods of Structuralization

Based on the examination items given by the physicians, the structuralized processing proposed in this paper focuses on two kinds of contents: 1) the summary content in the report, and 2) the results of special inspection items. The summary content includes three paragraphs in a report: the diagnosis, the comment/narrative, and the summary paragraphs. The results of special inspection items consist of five paragraphs: the paragraphs of EM, specimen type, gross description, LM, and DIF.

<1> Match dictionary to extract keyword list

Each paragraph with the summary content in an examination report is inputted into the structuralization module. Entity extraction is performed by matching the words in a paragraph with the vocabularies in the dictionary according to the pre-defined matching priorities. Then the phrases which represent entity concepts are extracted from the examination report to generate the examination report's keyword phrases. The matching priorities are as follows:

(1) The vocabularies with length less than 2 in the dictionary is not matched.

(2) The compound vocabularies consisting of more words have higher priorities to be matched.

(3) If the vocabularies have the same number of words, the vocabularies with higher frequency have higher priorities to be matched.

(4) Finally, the negative word list, as shown in Table 9 (in the Supplementary Files), is matched.

After completing the matching, the extracted phrases are combined sequentially according to some syntactic rules to generate a narrative short sentence with sufficient semantics. The syntactic rules for combination and the corresponding examples are shown in Table 10 (in the Supplementary Files). When a negative word appears, it is combined with its following keyword. After completing the above processing, a list *KP* of keyword phrases for the paragraph of the summary content is created.

<2> Categorize the summary content

The keyword phrases in the summary content need to be divided into different categories: procedure, primary diagnosis, and additional features. The vocabularies with the same base word are discovered when constructing the vocabulary dictionary. Accordingly, by inputting the base keywords of a certain procedure/disease as shown in Table 11 (in the Supplementary Files), the system can automatically match all the specific procedures and disease subtypes with the given base keywords.

Let $S_{procedure}$ and $S_{diagnosis}$ correspond to the sets of base keywords of procedure and diseases, respectively. By matching the base word of each keyword phrase in *KP* with the base keywords in $S_{procedure}$ and $S_{diagnosis}$, individually, the matched keyword phrases will be assigned to the corresponding categories. The rest of the keyword phrases, whose based words are not matched the base keywords of the procedure or the primary diagnosis, are categorized into the additional features.

□Example 3-3□

Assume the content in the diagnosis paragraph of an examination report is shown in Table 12 (in the Supplementary Files). The extracted phrases by matching with the vocabularies in the dictionary for the list of keyword phrases and denoted as *KP*, which consists of 7 phrases identified by 1 to 7. The base word of k1 matched the base keyword 'biopsy' of procedure. Therefore, keyword phrase 1 is assigned into the procedure category. By performing the similar processing, keyword phrases 2, 3, 6, and 7 are assigned into the primary diagnosis category because their base words 'glomerulopathy', 'glomerulosclerosis', 'nephropathy', and 'glomerulosclerosis' belong to the base keywords of main diagnosis. Finally, the rest of the keyword phrases 4 and 5 are categorized into the additional features.

Table 13 (in the Supplementary Files) shows an example with the results of special inspection items, which is a paragraph of DIF. For the results of special inspection items, the structuralized result must show the detailed observations for specific inspection items. The detailed results are usually described by continuous adjectives, such as "diffuse segmental coarse granular" appearing in the example. Therefore, in addition to the noun dictionary, an adjective dictionary is also used for matching when extracting the keyword phrase list *KP* from the results of special inspection items.

Next, according to the given specific keywords of each examination item for the structured report, the words in the extracted keyword phrase in *KP* are stemmed and compared with keywords of each examination item to decide which options of the examination item appearing. Negative words should be considered to decide 'present' or 'absence' of an examination item, as shown in Table 13.

3.2.2 Automatic Keyword Extraction for Special Inspection Items

Due to the large number of inspection items, it is cumbersome and time consuming for physicians to enumerate. Moreover, some items or their options may appear in the examination report but are not listed. Therefore, we proposed a probabilistic topic modeling method to automatically extract the candidate keywords of the examination items from the extracted keyword phrase list. The extracted candidate keywords for the inspection items can be provided to the physicians for further verification in order to reduce the effort of keyword enumeration manually.

For the five paragraphs with the results of special inspection items: EM, specimen type, gross description, LM, and DIF, the following keyword extraction processing is performed on each paragraph individually.

<1> Noise removal for the paragraph dictionary

The dictionary consists of two kinds of compound nouns, including the compound nouns of combining adjective and noun or the ones of combining continuous nouns.

We use the Lift measure to estimate the association degree between the singular nouns composing a compound noun p_l, as shown in Eq. 8.

$$Lift\left(p_l.w_n, p_l.w_m\right) = \frac{F(p_l.w_n \cap p_l.w_m)}{F\left(p_l.w_n\right) * F(p_l.w_m)} \text{ (Eq. 8)}$$

If the compound noun p_l is in the form of combining an adjective and nouns, let $p_l.w_n$ denote the adjective word in p_l and $p_l.w_m$ denote the following noun phrase in p_l. Otherwise, p_l is in the form of combining continuous nouns. Then let $p_l.w_n$ denote the first noun in p_l and and $p_l.w_m$ denote the following noun phrase in p_l. $F\left(p_l.w_n \cap p_l.w_m\right)$ denotes the number of times $p_l.w_n$ and $p_l.w_m$ cooccur. Besides, $F\left(p_l.w_n\right)$ and $F(p_l.w_m)$ denote the number of occurrences of $p_l.w_n$ and $p_l.w_m$,respectively.

For a compound noun p_l whose Lift value is greater than or equal to a given threshold value θ, it will be retained. Otherwise, it will be considered as a noisy vocabulary and removed from the dictionary.

<2> Find general adjectives

Among the compound nouns in the form of combining adjective and noun, we compute Entropy of each adjective to find general adjectives. Let JJ_n denote an adjective, and $t_1, \ldots t_m$ denote *m* different nouns appearing after JJ_n. Besides, $P\left(JJ_n + t_i \mid JJ_n\right)$ denotes the probability of t_i appearing after JJ_n.

$$Entropy\left(JJ_n\right) = -\int_{l=1}^{n} P\left(JJ_n + t_i \mid JJ_n\right) * \log P\left(JJ_n + t_i\right) \text{ (Eq. 9)}$$

According to the result of Eq. 9, if $Entropy\left(JJ_n\right)$ is larger than the threshold value 1, JJ_n is added into the list of general adjectives. A list of general adjectives is created after completing the entropy computation for all the adjectives in the constructed dictionary.

<3> Create candidate keyword list for examination items

We applied the LDA topic modeling method [3] to analyze the compound keyword phrases in the paragraph of certain special inspection items from the whole database for extracting the candidate keywords of the examination items. The LDA modeling assumes that each document is a mixture of various topics and each topic is described by a number of different hidden topic words. Accordingly, the LDA modeling aims to find a Dirichlet distribution which most fit the word observation of the documents in the database. Finally, we can get the probabilities of each document belonging to various hidden topics and the distributions of topic words for each hidden topic. For each paragraph of certain special inspection item, we collect the compound keyword phrases extracted from the corresponding paragraph in a report as the words in a document. After performing LDA topic modeling on the same paragraph for all the reports, the topic words with higher probabilities for each hidden topic is extracted as candidate keywords of the examination items.

For each paragraph with the results of special inspection items in a report, the sentences are parsed to get their POS tagging. The words with POS tags labeled as conjunctions, articles, pronouns, adverbs, auxiliary verbs, prepositions are removed, as shown in Figure 5. Then, the remaining words are compared with the vocabularies in the corresponding dictionary of the paragraph. The matched phrases are extracted to be the content in a document for LDA topic modeling, as shown in Figure 6. According to a given number num_T of topics, the result of LDA topic modeling will provide the topic words and their probabilities for each topic. The topic words with the top *k* highest probabilities for each topic are chosen. Let $Topic_n$ denote the *n*th topic and p_l denote a topic word of $Topic_n$. Besides, $AvgPTopic_n$ denotes the average probability of the top *k* highest probabilities for $Topic_n$ and $P_n(p_l)$ denotes the probability of p_l appearing in topic $Topic_n$. If $P_n(p_l) \geq AvgPTopic_n$, p_l is selected into the candidate keywords of the examination items as shown in Figure 7. After processing the topic words for each topic, the selected candidate keywords from each topic are collected into a set to provide the candidate keywords of the examination items in the paragraph.

<4> Extension for the candidate keyword list

In the result of the LDA topic modeling, the extracted candidate keyword list may miss the adjectives because they do not appear consecutively with the noun keywords in the report. Accordingly, it results in getting the incomplete options for the examination items. Therefore, the goal of this part of processing is to further find the general adjective to be the keywords for the options of examination items.

Let p_l denote a phrase in the candidate keyword list. For each adjective JJ_n in the general adjective list, if $JJ_n + p_l$ forms a compound noun phrase in the dictionary, $JJ_n + p_l$ is inserted into the candidate keyword list of examination items.

【Example 3-4】

Assume that a keyword candidate list, a general adjective list, and a dictionary of the compound nouns are given as shown in Table 13. Because the phrase '*basement membrane*' combined with the general adjective '*thick*' appearing in the dictionary, '*thick basement membrane*' is inserted into the candidate keyword list, where '*thick*' is a keyword of options for the observation item '*basement membrane*'. Similarly, '*diffuse*' is a keyword of options for'*foot processes effacement*' and '*mesangial*' is a keyword of options for'*expansion*'.

Results And Discussion

Three parts of experiments are performed to evaluate the proposed methods. The first part evaluates the effectiveness of eliminating typos by using the LCS (Longest Common Subsequence) method after establishing the medical dictionary. The second part computes the accuracy of the extracted examination item terms by adjusting the parameter settings in the calculation formula. Finally, the quality of the structured examination report constructed automatically is evaluated.

4.1 Experimental Data Source

The data set used in the experiments includes 476 examination reports of the patients in the Division of Nephrology of a University Hospital. All the contents in the report are in English. At first, the examination reports are first automatically segmented into the following eight paragraphs: (1) Main diagnosis (Diagnosis), (2) Electron Microscopy Examination (EM), (3) Examination Status of Electron Microscope (Comment/Narrative), (4) The Size and Condition of sliced Specimen (Specimen type), (5) Description of spliced Specimen (Gross description), (6) Examination of Light Microscope (LM), (7) Chromosome examination (DIF), (8) Report Conclusion (Summary). However, not all of the examination report contains all the eight paragraphs.

4.2 Evaluation on Typo Elimination using LCS

4.2.1 Evaluation Method

In subsection 3.2, a method is proposed for typos correction. In paragraph (1), (6), and (8) of the report, which contain more text descriptions and the typos occur frequently. Accordingly, the noun dictionaries are established for the three paragraphs, separately, and a noun dictionary for the whole corpus of the examination reports is constructed. For each of the four different dictionaries, 40 correction cases are randomly selected, which are manually labelled to decide whether the typos are correctly eliminated by using the LCS matching. Then a precision value is computed to show the effectiveness of the typos elimination strategy for each paragraph. Then, the precision achieved by performing on three paragraphs is averaged. After determining the appropriate 1/d value setting for typos correction by using the LCS matching, the Precision, Recall, and F1-score values of the typos correction by using the noun dictionary of the whole corpus is then computed.

4.2.2 Result of Experiments

The precision of typos correction by correcting the base word and the content word is shown in Figure 8 and Figure 9, respectively. The result shows that when 1/d is set to 1/3, the accuracy is poor, which is intuitive. When 1/d is set to 1/3, it means that a typo is allowed among every three letters. Accordingly, it causes two different correct words wrongly matched into the same word. Furthermore, the noun dictionary of the whole corpus performs more poor compared to the other three paragraph dictionaries. The result is not surprising because the corpus dictionary has more vocabulary than the paragraph dictionary, which causes it is easy to get an error typos correction by the LCS matching. As shown in Table 14 (in the Supplementary Files), most typos on the base word can be changed to the correct word (at least 90%) when the threshold is set to be 1/5. Table 14 and 15 shows some cases of the correct and wrong typo corrections, respectively. The wrong cases occur when the two correct words have many common characters in the same order as shown in the example of Table 15 (in the Supplementary Files).

The Precision, Recall, and F1-score values of typos correction by the three paragraph dictionary are shown in Figure 10. The result shows that the precision of typos correction performed on the paragraph (1) is the best. Although the performance on the paragraph (6) is worse, its recall achieves the best. The reason is because the number of distinct words appearing in paragraph (1) is lesser than the other paragraphs, so it is less likely to cause wrong matching of typos. On the other hand, on paragraph (6), the LCS method detects more typos than the actual number of typos, so that a higher Recall value is achieved. Overall, in terms of F1-score, the performance of typos corrections on these three paragraphs can achieve 0.62 or higher.

4.3 Evaluation on Keyword Extraction of Examination Items

4.3.1 Evaluate Method

Figure 11 shows the proposed methods for extracting the keywords introduced in subsection 4.2. We consider the following four processing components which may influence the extraction effectiveness: 1) dictionary construction, 2) the threshold value of LIFT filtering, 3) the topic number of LDA, and 3) whether to perform the keyword extension step. The corresponding experiments are from [Exp. 2-1] to [Exp. 2-4].

In this part of experiment, the data set consists of the paragraph (2), paragraph (6), and paragraph (7) in the reports, which describe the special examination items. The terms of the examination items in a structured form of the report, which are listed by a medical expert, are used as the correct answer. An extracted key phrase that contains a correct answer is considered to be predicted correctly. Accordingly, Precision, Recall, and F1-score are measured for the extracted keywords for each paragraph.

4.3.2 Experiment Result

□Exp. 2-1□Evaluation of keyword extraction based on various dictionary constructing strategies.

This experiment compares three different strategies for constructing the noun dictionaries: (1) nouns or compound nouns (labeled as NN+NN) that do not contain adjectives, (2) compound nouns (labeled as JJ+NN) that contain adjectives, (3) Union the compound nouns in the former two sets.

The precision, recall, and F1_score of the extracted keywords by using the dictionaries constructed by the three different methods are shown in Figure 12, Figure 13, and Figure 14, respectively.

The results show that the dictionary established by NN+NN is not as good as the others. The main reason is that the keywords composing an examination item usually contain adjective terms. Accordingly, the dictionary constructed by the discovered NN+NN phrases are not all correct answers and incomplete. By using the dictionary constructed from the union of the NN+NN and JJ+NN phrases, the recall of the extracted keywords is significantly higher than using the other two dictionaries. However, it results in a lower precision than using the (JJ + NN) dictionary. Overall, according to the average F1-score of the three paragraphs, the dictionary composing of the (NN+NN) phrases merged with the (JJ+NN) phrases is selected as the basis for extracting the examination keywords.

【Exp. 2-2】Evaluation of keyword extraction by changing the threshold setting of Lift measure

In this experiment, the threshold values θ of Lift measure for detecting the noun phrases is varied. The precision, recall, and F1_score of the extracted keywords by setting the various θ values are shown in Figure 15, Figure 16, and Figure 17, respectively. The results show that, when θ is set to be 0.2, all the paragraphs achieve the best performance for the evaluation metric. Accordingly, in the following experiments, θ is set to be 0.2.

【Exp. 2-3】Evaluation of keyword extraction by changing the number of LDA topics

In this experiment, by varying the number of LDA topics for detecting the noun phrases, the precision, recall, and F1_score of the extracted keywords are measured. The results are shown in Figure 18, Figure 19, and Figure 20, respectively.

According to the results of experiments, when num_T is set to be 10, the highest precision of the extracted candidate keyword list can be obtained. Besides, when num_T is set to be 30, the highest recall can be obtained. For the DIF paragraph, numbered 7, the recall is even higher than 0.9. The reason for getting the above results is that the more the topic number is, the more keywords will be extracted by the LDA topic modeling. Accordingly, more keywords of the examination items are discovered. However, more keywords also get more divergent keywords, which cause accuracy of the extracted keywords drop a little. Overall, when num_T is 10, the superior performance of F1-score can be achieved.

【Exp. 2-4】 Evaluation on the extension for the candidate keywords list.

According to the proposed method for expanding the candidate keyword list proposed in section 4.2, the purpose of this experiment aims to evaluate whether the method improves the effectiveness of the extracted keyword list. The results of precision, recall, and F1 measure by comparing before and after performing the proposed method are shown in Figure 21. It shows that the proposed method indeed effectively improves the precision and recall values of the extracted candidate keyword list for each paragraph. In other words, to find the corresponding general adjective vocabulary provides a more correct and complete list of keywords for the examination items. Table 16 (in the Supplementary Files) shows the words which are not extracted by the proposed method, most of these words have low frequency of occurrence in the reports. The other lost keywords whose probability for a specific topic is lower than the average of the top k probabilities for that topic. It implies these observations rarely occur in the reports of patients.

4.4Evaluation on Structuralization of the Summary Content

4.4.1 Evaluate Method

According to the method proposed in subsection 4.1, this experiment takes the structuralization result for paragraphs (1) Main diagnosis, (3) Comment/Narrative, and (8) Summary for evaluation. The experts manually mark whether the extracted keyword phrases provided important medical information. If the extracted phrase has some missing word not extracted from the report, it is considered to be an incorrect result. In this experiment, 50 structured results of the reports were randomly sampled for each of the three paragraphs, the Precision metric is computed to show the performance according to the manually labeled results.

4.4.2 Experiment Result

The result of this experiment is shown in Figure 22. It can be found that, for these abstract paragraphs, the extracted diagnoses and observations can achieve a precision of at least 0.9. For the Comment/Narrative paragraph, it achieves precision of 0.98. It implies that the proposed method can well provide a structured form for most of the examination reports for the abstract paragraphs. The small number of errors, as shown Table 17 (in the Supplementary Files), is mainly caused by the missing keywords which are not completely extracted in the keyword phrases. The reason is the corresponding phrases do not satisfy the combination rules of keyword phrase extraction, but the POS tag patterns rarely appear.

Conclusion And Future Work

5.1 Conclusion

This study aims to automatically generate a structured form for a textual examination report. At first, based on the part-of-speech tagging results, some patterns are designed to extract candidates of medical vocabulary from the corpus of nephrology examination reports. Besides, the possible typos and clinical meaningless words are filtered out to construct a medical vocabulary dictionary for each examination paragraph. For the paragraphs of special examination items, the noisy words are removed from the phrases in the established dictionaries. Then the LDA topic modeling is applied to extract the candidate keyword phrases of the examination items. For the abstract paragraphs, the content in each paragraph is matched with the vocabulary dictionary to extract keyword phrases, which are then merged into complete medical terms and assigned in to different categories to show a structured form of the paragraph. The results of a series of experiments show that the methods proposed in this paper can effectively construct a structural form of examination reports. Furthermore, the keywords of the popular examination items can be extracted correctly. The above techniques will help automatic processing and analysis of medical textural reports.

5.2 Future Work

According to the results of experiments, in the structured form of the paragraphs, a small number of keywords are missing which are not extracted by the currently proposed part-of-speech pattern rules. We will consider to collect a larger database of examination reports, combined with a pre-constructed medical vocabulary, to automatically learn syntactic pattern to provide a more complete and effective extraction method for keyword phrases. Furthermore, how to analyze the cause of main diagnose from the keywords of the examination item will be studied in the future.

Abbreviations

EMR - electronic medical records

EM – electron microscope

LM – light microscope

LDA – latent Dirichlet allocation

NLP – natural language processing

PAD – peripheral arterial disease

CRFs – conditional random fields

SVM – support vector machine

LCS – longest common subsequence

POS Tagging – part of speech tagging

NN – noun

JJ – adjective

VBD – past tense verb

CD – quantifier

VBN – past participle

Declarations

Ethics approval and consent to participate

The study was approved by the Big Data Center of CMUH and the Research Ethical Committee/Institutional Review Board of CMUH (CMUH105-REC3-068).

Consents were waived by the approval of the Research Ethical Committee/Institutional Review Board of CMUH (CMUH105-REC3-068).

- Consent for publication

All authors have read and approved the submission of this manuscript.

- Availability of data and material

Authors agreed to make the relevant anonymised patient level data available on reasonable request.

- Competing interests

All authors have nothing to disclose.

- Funding

This study was supported by China Medical University Hospital (grant number: DMR-100-004) and the Ministry of Science and Technology of Taiwan (grant number: 106-2314-B-039-041- MY3)

- Authors' contributions

CWT and CCK designed the study, PHW, JLK and ALPC conducted data extraction strategies and performance analysis. PHW and AY drafted the manuscript. JLK and ALPC critically revised the manuscript. CWT, CCK, PHW, and JLK participated in the literature search, data preparation, and manuscript editing.

- Acknowledgements:

Not applicable.

References

[1] A. R. Aronson,"Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001.

[2] S. Balaneshin-kordan, A. Kotov, and R. Xisto. Wsu-Ir at trec 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In Proc. of the Text

REtreival Conference (TREC), 2015.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In Proc. of the Journal of Machine Learning Research (JMLR), 2003.

[4]R. Feldman, O. Netzer, A. Peretz, and B. Rosenfeld. Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. In Proc. of Knowledge Discovery and Data Mining (KDD), 2015.

[5] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding physiological state: mortality modelling in intensive care units. In Proc. of the Knowledge Discovery and Data Mining (KDD), 2014.

[6] T. R. Goodwin, and S. M. Harabagiu. Medical question answering for clinical decision support. In Proc. of the International Conference on Information and Knowledge Management (CIKM), 2016.

[7] Y. Jo, N. Loghmanpour, and C. P. Rose. Time series analysis of nursing notes for mortality prediction via a state transition topic model. In Proc. of the International Conference on Information and Knowledge Management (CIKM), 2015.

[8] J. Lafferty, A. McCallum, and F. CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data.", 2001.

[9] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In Proc. of the 2010 workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, 2010.

[10]L.-W. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In Proc. of the American Medical Informatics Association (AMIA), 2012.

[11] X. Liu, and H. Chen. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In Proc. of the *International Conference on Smart Health* (ICSH), 2013.

[12] N. Nandhakumar, et al. "Clinically Significant Information Extraction from Radiology Reports." Proceedings of the 2017 ACM Symposium on Document Engineering. ACM, 2017.

[13] M. Paterson, and V. Dančík. Longest common subsequences. In Proc. of the Mathematical Foundations of Computer Science (MFCS), 1994.

[14] X. Rong, Z. Chen, Q. Mei, and E. Adar. EgoSet: exploiting word ego-networks and user-generated ontology for multifaceted set expansion. In Proc. of the International Conference on Web Search and

Data Mining (WSDM), 2016.

[15]G. K. Savova, et al. "Discovering peripheral arterial disease cases from radiology notes using natural language processing." AMIA Annual Symposium Proceedings. Vol. 2010. American Medical Informatics Association, 2010.

[16] G. K. Savova, et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." Journal of the American Medical Informatics Association 17.5 (2010): 507-513.

[17] S. Sethi, et al. Mayo clinic/renal pathology society consensus report on pathologic classification, diagnosis, and reporting of GN. In Proc. of the Journal of the American Society of Nephrology (JASN), 2015.

[18] Stanford CoreNLP – Core natural language software *https://stanfordnlp.github.io/CoreNLP.*

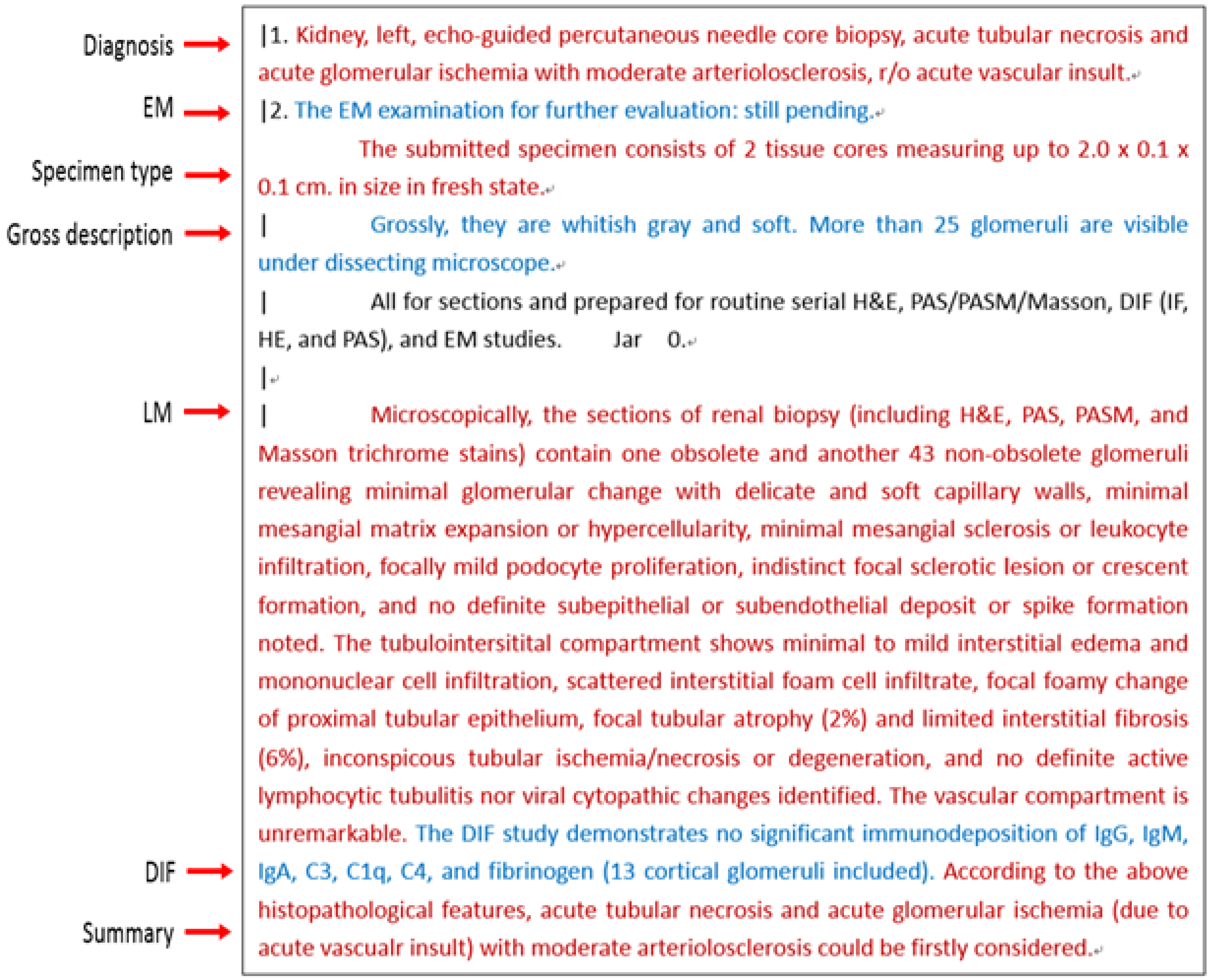
[19] N. Tandon, G. D. Melo, A. De, and G. Wrikum. Knowlywood: mining activity knowledge from hollywood narratives. In Proc. of the International Conference on Information and Knowledge Management (CIKM), 2015.

[20]Y. Wang, et al. "Clinical information extraction applications: a literature review." Journal of biomedical informatics 77 (2018): 34-49.

Tables

Due to technical limitations, tables are only available as downloads in the supplemental files section.

Figures


Diagnosis
EM
Specimen type
Gross description
LM
DIF
Summary

|1. Kidney, left, echo-guided percutaneous needle core biopsy, acute tubular necrosis and acute glomerular ischemia with moderate arteriolosclerosis, r/o acute vascular insult.

|2. The EM examination for further evaluation: still pending.

The submitted specimen consists of 2 tissue cores measuring up to 2.0 x 0.1 x 0.1 cm. in size in fresh state.

| Grossly, they are whitish gray and soft. More than 25 glomeruli are visible under dissecting microscope.

| All for sections and prepared for routine serial H&E, PAS/PASM/Masson, DIF (IF, HE, and PAS), and EM studies. Jar 0.

|

| Microscopically, the sections of renal biopsy (including H&E, PAS, PASM, and Masson trichrome stains) contain one obsolete and another 43 non-obsolete glomeruli revealing minimal glomerular change with delicate and soft capillary walls, minimal mesangial matrix expansion or hypercellularity, minimal mesangial sclerosis or leukocyte infiltration, focally mild podocyte proliferation, indistinct focal sclerotic lesion or crescent formation, and no definite subepithelial or subendothelial deposit or spike formation noted. The tubulointersititial compartment shows minimal to mild interstitial edema and mononuclear cell infiltration, scattered interstitial foam cell infiltrate, focal foamy change of proximal tubular epithelium, focal tubular atrophy (2%) and limited interstitial fibrosis (6%), inconspicous tubular ischemia/necrosis or degeneration, and no definite active lymphocytic tubulitis nor viral cytopathic changes identified. The vascular compartment is unremarkable. The DIF study demonstrates no significant immunodeposition of IgG, IgM, IgA, C3, C1q, C4, and fibrinogen (13 cortical glomeruli included). According to the above histopathological features, acute tubular necrosis and acute glomerular ischemia (due to acute vascualr insult) with moderate arteriolosclerosis could be firstly considered.

Figure 1

Example of a nephrology examination report.

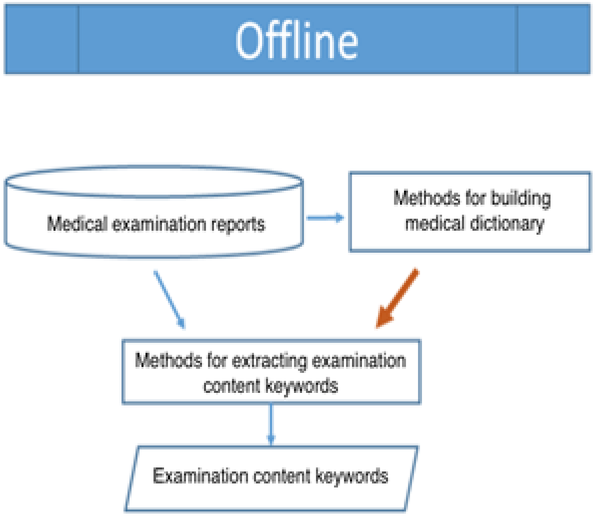

Offline
Medical examination reports
Methods for building medical dictionary
Methods for extracting examination content keywords
Examination content keywords

Figure 2

Flowchart of the offline processing.

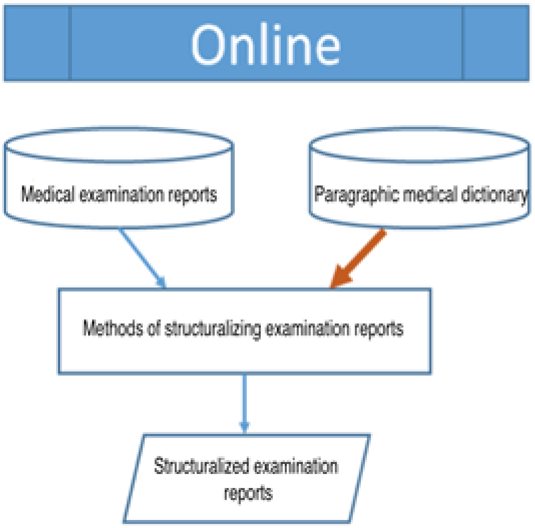
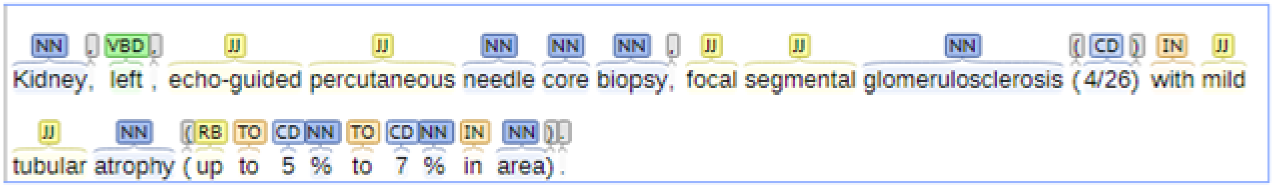

Online
Medical examination reports
Paragraphic medical dictionary
Methods of structuralizing examination reports
Structuralized examination reports

Figure 3

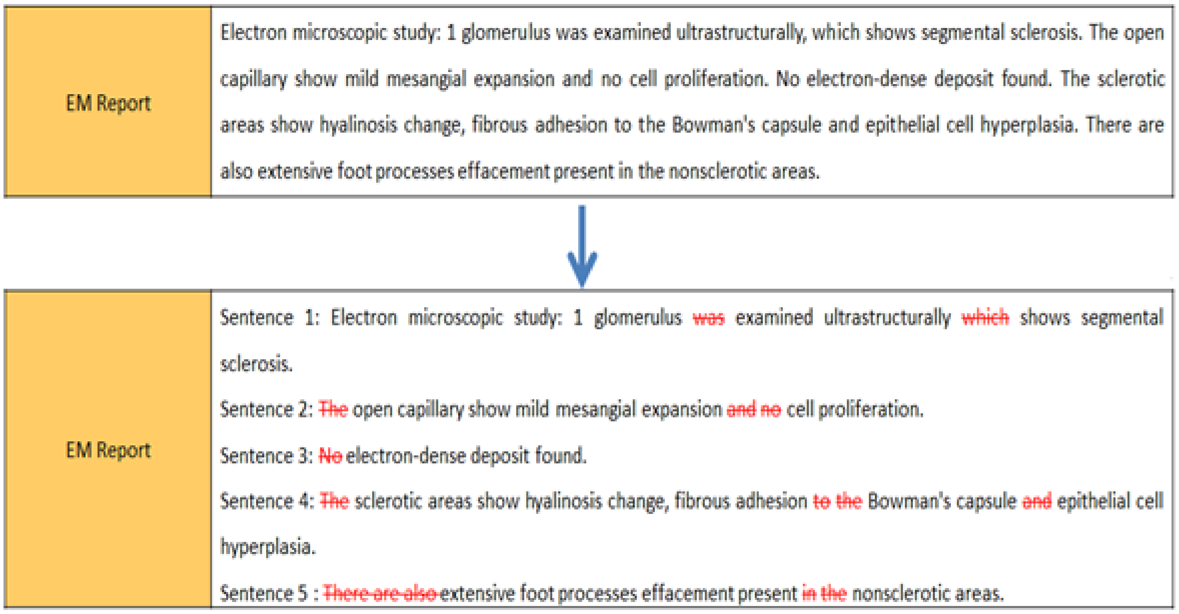
Flowchart of the online processing.


NN , VBD , JJ JJ NN NN NN , JJ JJ NN (CD) IN JJ
JJ NN (RB TO CD NN TO CD NN IN NN) .

Kidney, left, echo-guided percutaneous needle core biopsy, focal segmental glomerulosclerosis (4/26) with mild tubular atrophy (up to 5 % to 7 % in area).

Figure 4

Example of the Part-of-Speech tagging result.

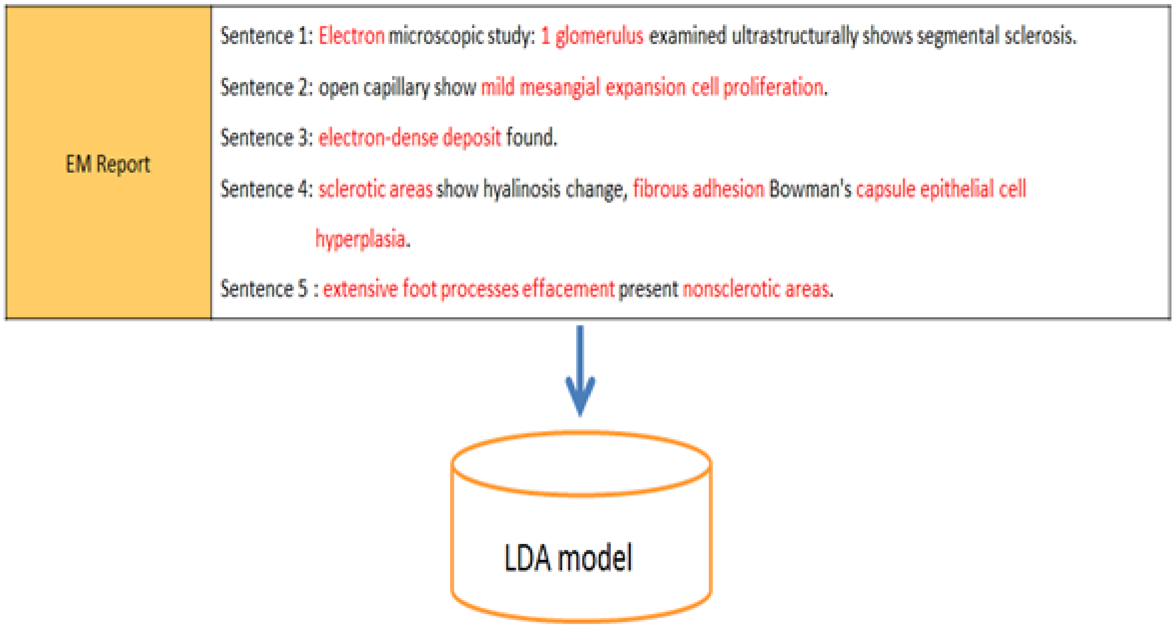


EM Report	Electron microscopic study: 1 glomerulus was examined ultrastructurally, which shows segmental sclerosis. The open capillary show mild mesangial expansion and no cell proliferation. No electron-dense deposit found. The sclerotic areas show hyalinosis change, fibrous adhesion to the Bowman's capsule and epithelial cell hyperplasia. There are also extensive foot processes effacement present in the nonsclerotic areas.

EM Report	Sentence 1: Electron microscopic study: 1 glomerulus ~~was~~ examined ultrastructurally ~~which~~ shows segmental sclerosis. Sentence 2: ~~The~~ open capillary show mild mesangial expansion ~~and no~~ cell proliferation. Sentence 3: ~~No~~ electron-dense deposit found. Sentence 4: ~~The~~ sclerotic areas show hyalinosis change, fibrous adhesion ~~to the~~ Bowman's capsule ~~and~~ epithelial cell hyperplasia. Sentence 5 : ~~There are also~~ extensive foot processes effacement present ~~in the~~ nonsclerotic areas.

Figure 5

Example of sentence separation and semantic meaningless words removing.



EM Report

Sentence 1: Electron microscopic study: 1 glomerulus examined ultrastructurally shows segmental sclerosis.

Sentence 2: open capillary show mild mesangial expansion cell proliferation.

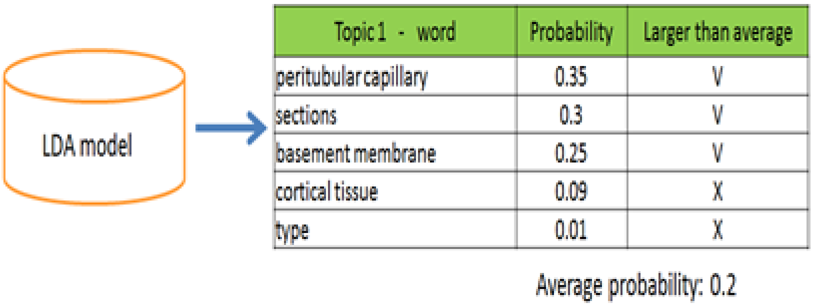
Sentence 3: electron-dense deposit found.

Sentence 4: sclerotic areas show hyalinosis change, fibrous adhesion Bowman's capsule epithelial cell hyperplasia.

Sentence 5 : extensive foot processes effacement present nonsclerotic areas.

Figure 6

The extracted keyword phrases from the paragraph dictionary.



Topic 1 - word	Probability	Larger than average
peritubular capillary	0.35	V
sections	0.3	V
basement membrane	0.25	V
cortical tissue	0.09	X
type	0.01	X

Average probability: 0.2

Figure 7

Example of selecting candidate keywords of the examination items.

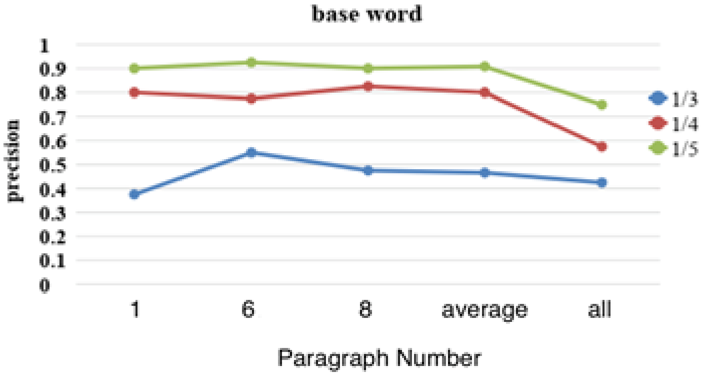


Figure 8

The precision of typos correction by using LCS to correct the base word.

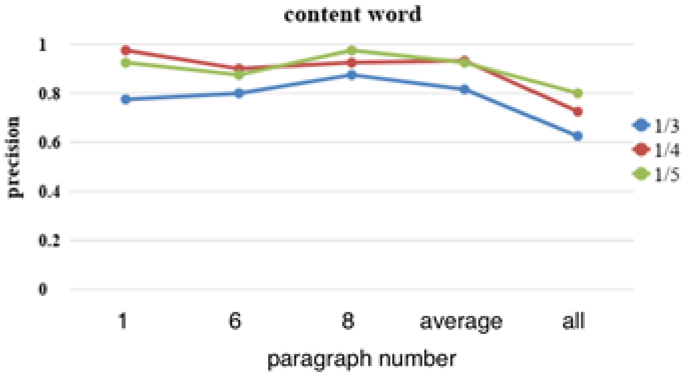


Figure 9

The precision of typos correction by using LCS to correct the content word.

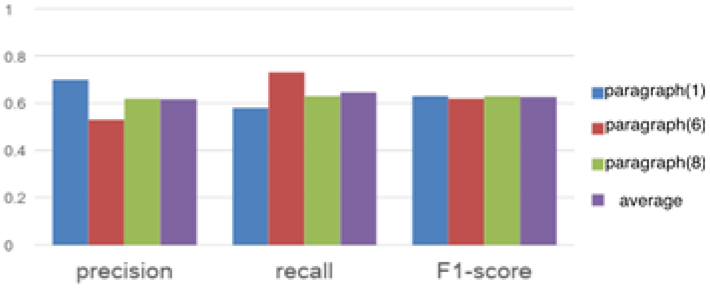


Figure 10

The result of Precision, Recall, and F1-score of typos correction by using paragraph dictionaries.

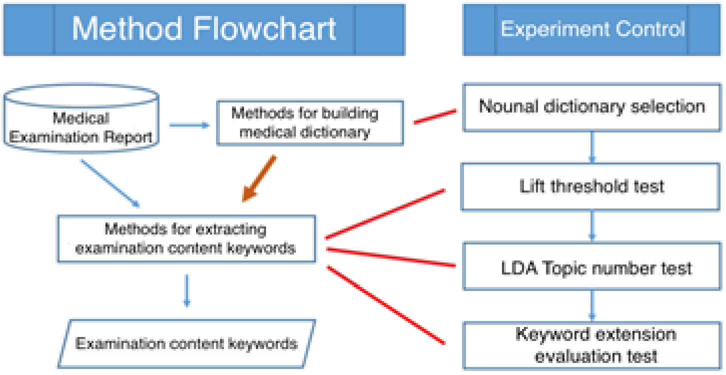

Method Flowchart
Experiment Control
Medical Examination Report
Methods for building medical dictionary
Methods for extracting examination content keywords
Examination content keywords
Nounal dictionary selection
Lift threshold test
LDA Topic number test
Keyword extension evaluation test

Figure 11

Method and Experiment Flowchart.

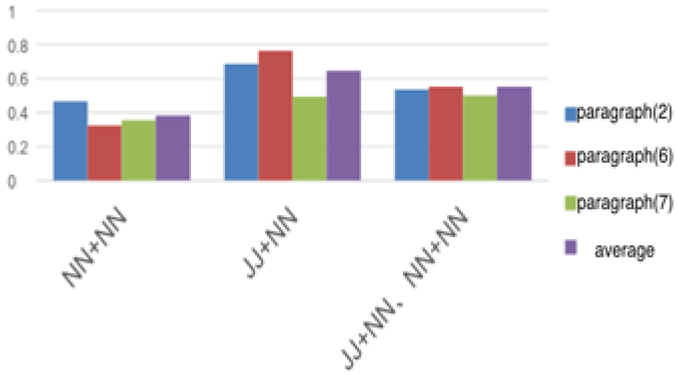


Figure 12

The precision of keyword extraction based on different methods of dictionary construction.

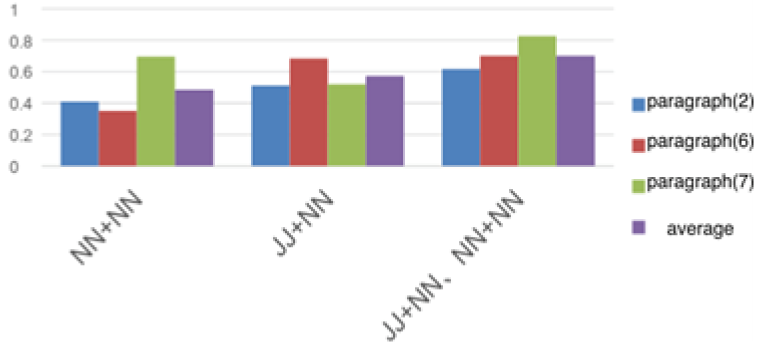


Figure 13

The recall of keyword extraction based on different methods of dictionary construction.

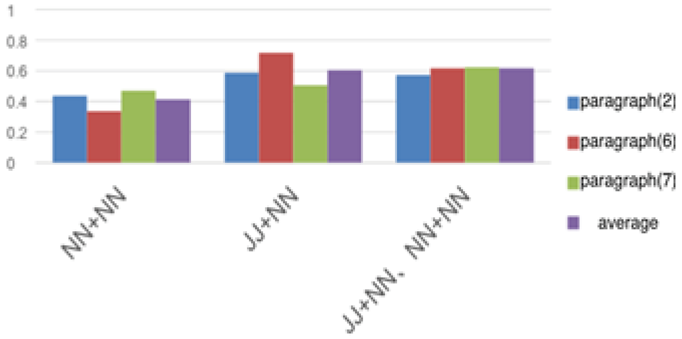


Figure 14

The F1-score of keyword extraction based on different methods of dictionary construction.

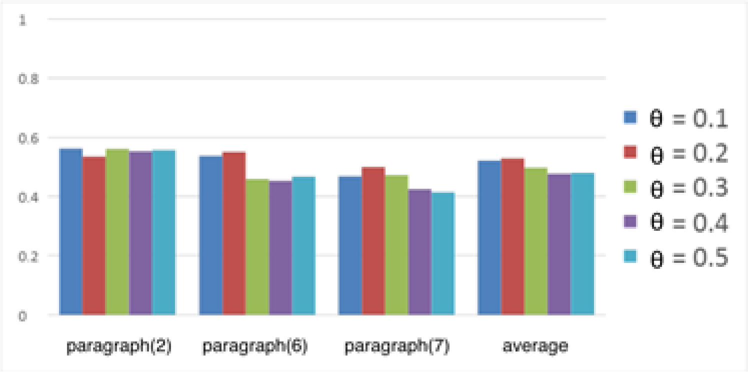


Figure 15

The precision of keyword extraction based on different threshold setting on Lift measure.

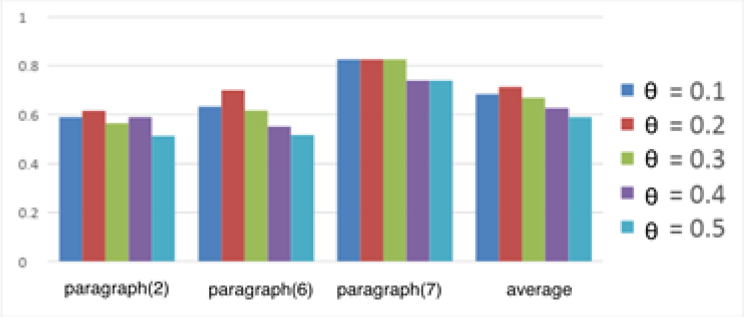


Figure 16

The recall of keyword extraction based on different threshold setting on Lift measure

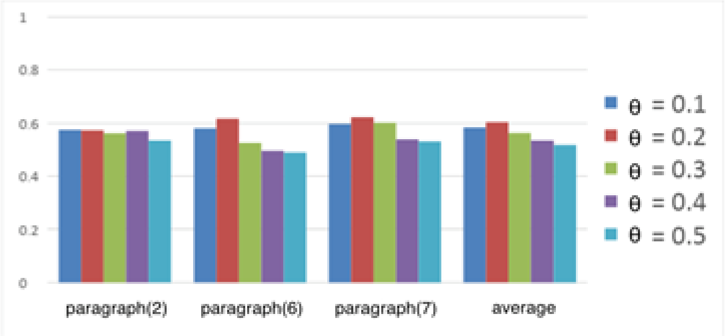


Figure 17

The F1-score of keyword extraction based on different threshold setting on Lift measure.

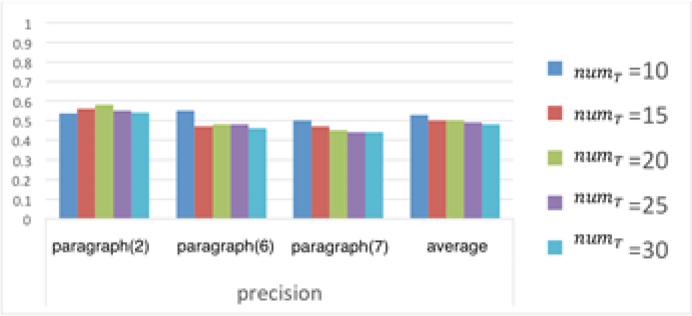


Figure 18

The precision of the extracted candidate keyword list by setting different LDA topic number.

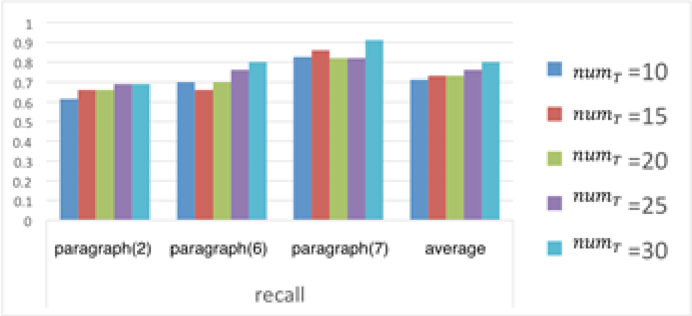


Figure 19

The recall of the extracted candidate keyword list by setting different LDA topic number.

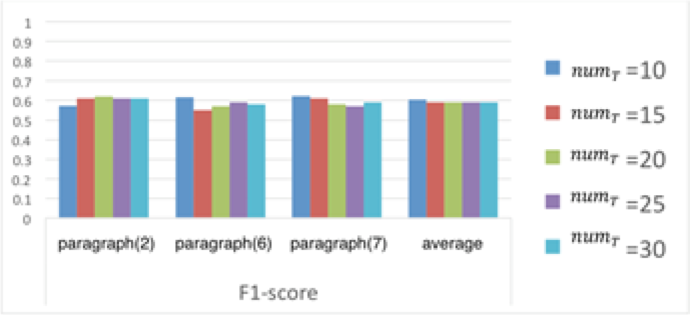


Figure 20

The F1-score of the extracted candidate keyword list by setting different LDA topic number.

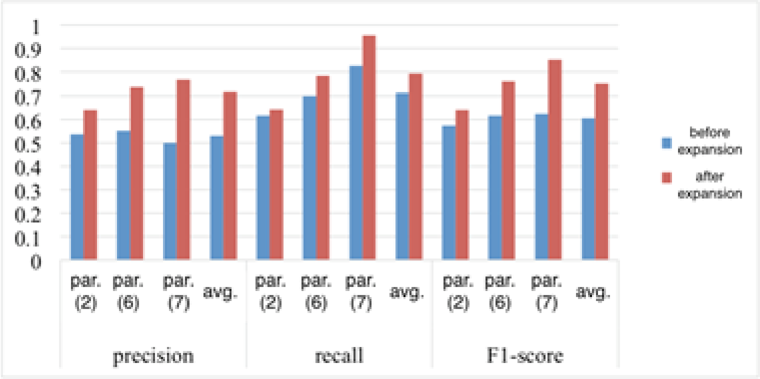


Figure 21

The performance of examination content keyword based on expanding keyword candidate list.

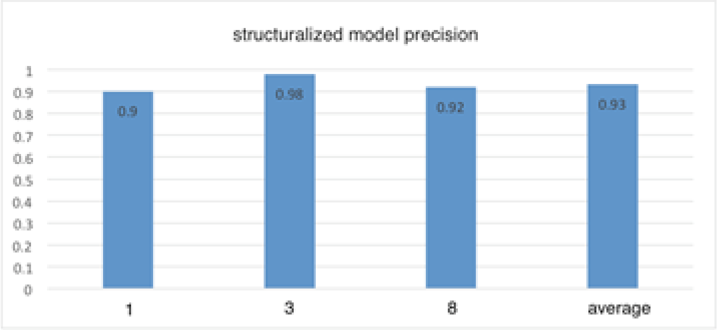


Figure 22

The precision of the structured form for the abstract paragraphs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.jpg
- supplement2.jpg
- supplement3.jpg
- supplement4.jpg
- supplement5.jpg
- supplement6.jpg
- supplement7.jpg
- supplement7.jpg
- supplement9.jpg
- supplement10.jpg
- supplement11.jpg
- supplement12.jpg
- supplement13.jpg
- supplement14.jpg
- supplement15.jpg
- supplement16.jpg
- supplement17.jpg
- supplement18.jpg