Structuring electronic dental records through deep learning for a clinical decision support system

Qingxiao Chen (iD)
Peking University School and Hospital of Stomatology; National Engineering Laboratory for Digital and Material
Technology of Stomatology; Peking University; National Clinical Research Center for Oral Diseases, PR China

Xuesi Zhou
Ji Wu
Tsinghua University, PR China

Yongsheng Zhou
Peking University School and Hospital of Stomatology, PR China

Abstract

Extracting information from unstructured clinical text is a fundamental and challenging task in medical informatics. Our study aims to construct a natural language processing (NLP) workflow to extract information from Chinese electronic dental records (EDRs) for clinical decision support systems (CDSSs). We extracted attributes, attribute values, and tooth positions based on an existing ontology from EDRs. A workflow integrating deep learning with keywords was constructed, in which vectors representing texts were unsupervised learned. Specifically, we implemented Sentence2vec to learn sentence vectors and Word2vec to learn word vectors. For attribute recognition, we calculated similarity values among sentence vectors and extracted attributes based on our selection strategy. For attribute value recognition, we expanded the keyword database by calculating similarity values among word vectors to select keywords. Performance of our workflow with the hybrid method was evaluated and compared with keyword-based method and deep learning method. In both attribute and value recognition, the hybrid method outperforms

Corresponding author:
Yongsheng Zhou, Department of Prosthodontics, Peking University School and Hospital of Stomatology, 22 Zhongguancun Avenue South, Haidian District, Beijing 100081, PR China.
Email: kqzhouysh@hsc.pku.edu.cn

the other two methods in achieving high precision (0.94, 0.94), recall (0.74, 0.82), and *F* score (0.83, 0.88). Our NLP workflow can efficiently structure narrative text from EDRs, providing accurate input information and a solid foundation for further data-based CDSSs.

Introduction

Electronic health records (EHRs) are official documents recorded by doctors which contain abundant medical information about patients including patients' individual desire, physical examinations, radiological examination results, and lab tests.[1–4] Medical information extraction from EHRs becomes a fundamental step for analyzing EHR data and constructing data-based models in many applications including hospital management, EHR template construction, decision support system research, etc. In our previous research, we constructed a clinical decision support system (CDSS) for removable partial denture design in dentistry.[5] As the denture design is associated with many determinants of the patient's oral conditions that are sufficiently described in the oral examination section (OES) of electronic dental records (EDRs), we intend to extract information from EDRs to build data-based models for CDSS. In this study we focus on constructing an efficient workflow to extract information from narrative dental records and instantiate them with our ontology.

Our primary goal is to transfer textual data into well-structured instances which has not been solved properly by researches. Schleyer et al.[6] constructed the Oral Health and Disease Ontology and extracted information from the tables of their EDR system. However, the OHD covers general knowledge in dentistry instead of the specific concepts to support the CDSS for denture design. And their method could not handle EDRs in narrative format, which is a major challenge for information extraction.[1,7,8] To deal with narrative EDRs, researchers have proposed and implemented some methods to deal with narrative dental records. Christensen et al.[9] used a sentence-level text analyzer ONYX, which is a semantic network based model, for semantic analyses of dental clinical text. Irwin et al.[10] described methods to develop and evaluate dental semantic representations for natural language processing (NLP). Both of the two researches extract partial information from EDRs involving a small fraction of entities, concepts, and relations, which makes the research less complex. And the rule-based or learning-based methods they proposed require intensive manual effort including summarizing expert rules and feature engineering. Considering the abundant information we intend to extract, it is labor-consuming to manually build rules that could cover all possible information and expressions.

In our study, we constructed a workflow to achieve information extraction (IE) from the OES shown in Figure 1. Our protocol is at first to read texts from the OES; then word segmentation is established by software Language Technology Platform (LTP).[11] The IE procedure comprises three tasks: attribute recognition, attribute value recognition, and tooth position recognition, corresponding to classifications of oral examination, relevant results, and tooth position, respectively. We compared the performance of Sentence2vec and bidirectional encoder representations from transformers (BERT) in our corpus and implemented Sentence2vec in attribute recognition.[12] We utilized Word2vec in attribute value recognition.[13–15] Vectors representing sentences and words are obtained during the training process of the Skip-gram model. After the IE process, attributes, relevant values, and tooth positions are instantiated into our ontological paradigm.
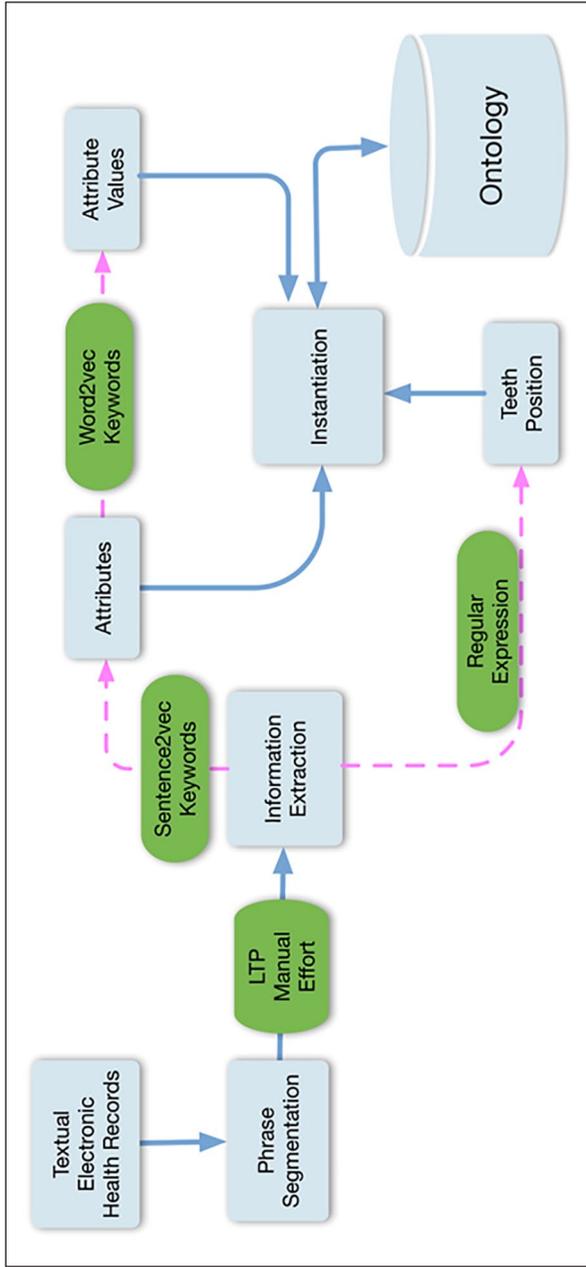
Figure 1. Framework of EDR structuring.

The entire workflow of EDR structuring is shown, in which blue squares represent every procedure and green squares represent methods we applied.

Methods

A piece of EDR in Peking University Hospital of Stomatology, electronically recorded by dentists in narrative formats, represents patients' dental visiting information during one appointment. Six major sections comprise one complete record, namely chief complaint, medical history, oral examination, diagnosis, treatment plan, and disposal. Among them, we utilized the text of the OES with detailed description of patients' oral conditions including physical examinations and imaging tests.

Our cohort involved 8000 de-identified EDRs from the database of Prosthodontics Department, Peking University Hospital of Stomatology. All private information was erased from the original records. According to our statistics, there are 10,492 different sentences, involving 76,327 words and 130,068 characters in total.

We annotated 2000 EDRs from the entire set. One dentist and two researchers designed the annotation guideline for attribute and value recognition (Appendix section); annotation in our study was labeled sentence by sentence. During annotation, the two researchers first annotated 200 records with an annotation program we developed. The dentist checked the annotations and discussed with the two researchers until three members reached agreements. Then the two researchers annotated the whole corpus separately. Cohen's κ between the two annotators is 0.872.[16]

Word segmentation

Words consisting of characters are the smallest units that could express semantic meanings in Chinese. Texts in the OES are semi-structured; Sentences are in a regular form as "tooth position, oral examination (corresponding to attribute), result (corresponding to attribute value)." For instance, "6/8765/4567/18剩余牙槽嵴中度吸收"(moderate resorption of residual ridge in tooth positions 6/8765/4567/18) could be segmented as "6/8765/4567/18," "剩余牙槽嵴," "中度吸收" (tooth position 6/8765/4567/18, residual ridge, medium resorption). We utilized the LTP tool for segmentation, which is produced by the Harbin Institute of Technology in China.[11]

Information extraction

We separated the whole IE task into three parts as recognition of attribute, attribute value, and tooth position. For attribute and value recognition, 80% of the 2000 annotated EDRs were randomly selected as the training set; the remaining were treated as the test set.

Attribute recognition

We designed a hybrid method integrating sentence vectors with keywords of attributes for recognition (Figure 2(a)). For every input sentence, a 100-dimensional vector was acquired through Sentence2vec, then used to calculate cosine similarity values with all sentences in the training set. A keyword-based method was also applied for attribute recognition in the same sentence. Thus, attribute recognition for a particular input sentence is obtained either by similarity calculation or by keyword matching.

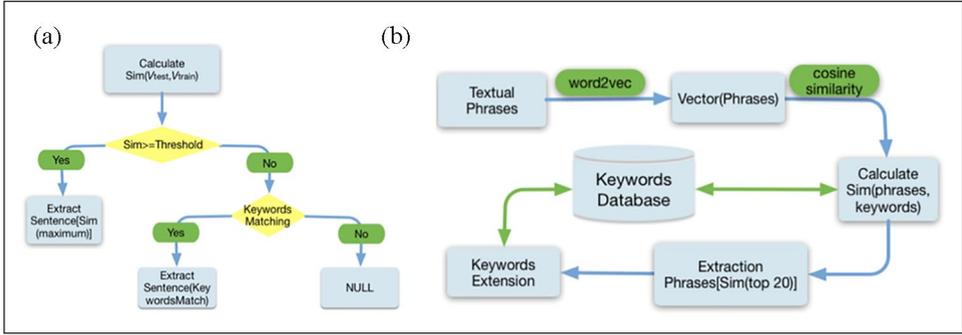The strategy to select attributes was as follows:

Figure 2. (a) Hybrid method of attribute recognition. Principles of attribute selection are described, using keyword-based and deep learning methods, and (b) workflow of attribute values recognition. Deep learning methodology was applied to generate new keywords to expand the keywords database.

(a) If similarity value is larger than the threshold, the attribute of the input sentence is recognized as the sentence attribute in the training set.

(b) If similarity value is less than the threshold and matches the keyword, the attribute of the input sentence is the keyword attribute.

(c) If similarity value is less than the threshold and keyword matching fails, then the output is null and no information is extracted.

During attribute recognition, the threshold is set to switch between keyword-based and deep learning methods. To select a threshold value, we plotted precision, recall, and F1 measure at finite and discretized threshold values, shown in Figure 3(a) and (b). At various thresholds, attribute recognition achieves different levels of performance. The strategy of threshold selection involves a balance between high precision and high recall. Here, we set the threshold to be 0.9.

Vector learning. We utilized Sentence2vec and BERT to train sentence vectors and applied Sentence2vec to better learn vectors of sentences from 8000 unlabeled EDRs.[13] Specifically, skip-gram model and hierarchical softmax optimization methods were implemented to learn 100-dimensional vectors from variable-length sentences. The skip-gram model was used to find word representations that could predict the neighboring words in a sentence.[14]

Given a series of training words, w_1, w_2, w_3, ... w_t,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p\left(w_{t+j}|w_t\right) \tag{1}$$

where c is the size of the training set, the objective of the skip-gram model was to maximize the log probability. After iteration, the largest probability value was calculated and word vectors were learned as parameters of the model. Sentence vectors are matrices involving word vectors with an additional sentence token in a sentence.

Similarity calculation. For attribute recognition, an input sentence, with a fixed-length vector, was compared with all sentences in the training set. It is acknowledged that semantically similar sentences exhibit similar vector representations in space. Therefore, cosine similarity was calculated
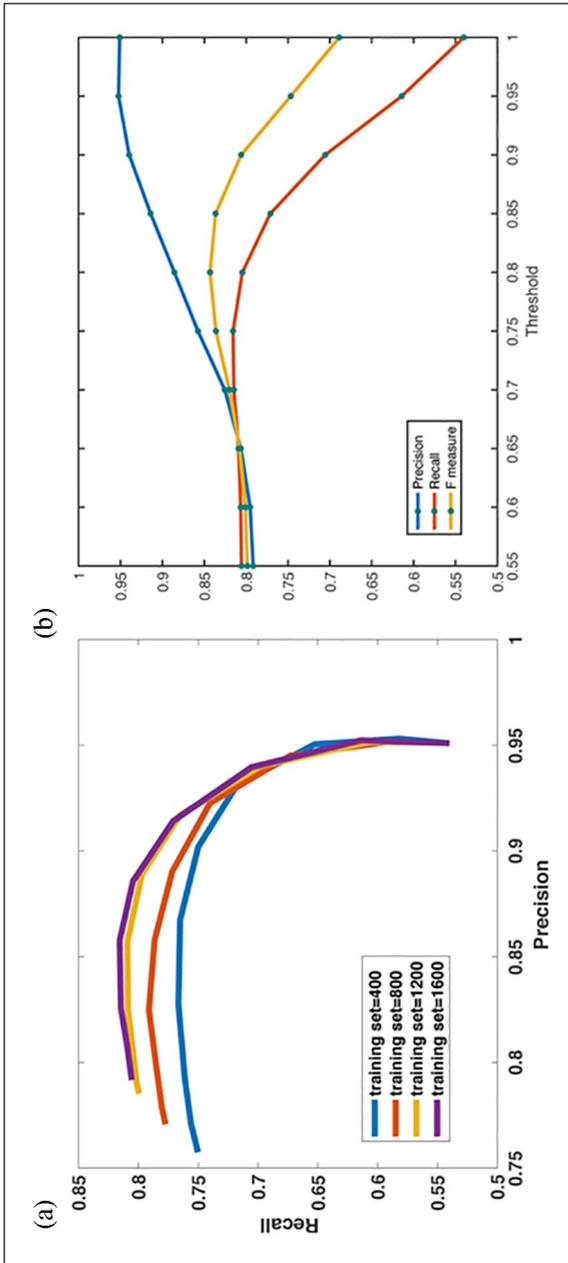
Figure 3. (a) Performance under different numbers of cases in the training set. The number of cases in the training set can influence the model's performance. As seen in this figure, a larger training set increases recall and precision, which helped us to determine the ideal size of the training set, and (b) performance comparison with different thresholds. In defining the principles of attribute selection, the threshold of similarity value is demanding; thus, we set the threshold to maintain overall performance of our model.

between the vector of the input sentence and all vectors in the training set and similar sentences could be distinguished by high similarity values.

Attribute value recognition

For attribute value recognition, we constructed the keyword database with 366 words and concepts and expanded it via learning word vectors (Figure 2(b)). One or two keywords were initially manual selected to represent each attribute value according to medical vocabularies and textbooks. Then, all words representing the attribute's values, in both the keyword dataset and the 8000 EHRs, were used to learn vectors, using the Skip-gram model.[14,15,17] A 200-dimensional vector was learned for each word.

The strategy of expanding keywords is as follows. For each sentence in the training set, cosine similarity was calculated among vectors of words with given keywords in the keyword dataset. For every attribute value, 20 words with the highest similar values were extracted and added as potential keywords into the keyword database. Two annotators reviewed the top 20 words for each attribute value and selected the final keywords. Eventually 473 keywords comprised the database.

Tooth position recognition

Tooth positions were described in numbers and symbols, which are clear for use in construction of rules. Regular expressions were built for tooth position recognition.

Instantiation

After information extraction, all extracted attributes, values, and teeth position were mapped into the previously-built ontological paradigm. The ontology represents knowledge of denture design in a structured and formalized way and keeps consistent with the existing CDS model.[5] Based on data properties and object properties embedded in the ontology, the mapped information forms instances representing oral health information of a patient's one appointment. To obtain instantiation, Class *appointment* was built to represent every patient appointment. Data properties *tooth ordinal* and *tooth zone* were added to represent tooth position. Object properties *left_first_tooth* and *right_first_tooth* were built to define boundaries of continuously missing teeth.

A relationship map of a patient's instance is illustrated in Appendix Figure 5. Here, Patient A defines the patient, Appointment 1 refers to this visit, and Oral Conditions references patient A's oral examinations. Specifically, Instances Tooth 26, Tooth 27, and Tooth 28 were constructed to describe examinations of teeth 26, 27, and 28 (FDI notation); Instance Edentulous Space was constructed to represent examination of edentulous spaces, and is connected with Instances Tooth 26, Tooth 27, and Tooth 28 by object properties *left_first_tooth* and *right_first_tooth*, indicating that missing teeth in the edentulous space are teeth 26, 27, and 28 (FDI notation).

Results

We performed the evaluation in the test set containing 400 annotated EDRs.[18] Precision (P), Recall (R), and F1 scores were calculated to quantitatively employ efficiency of the proposed methods where,

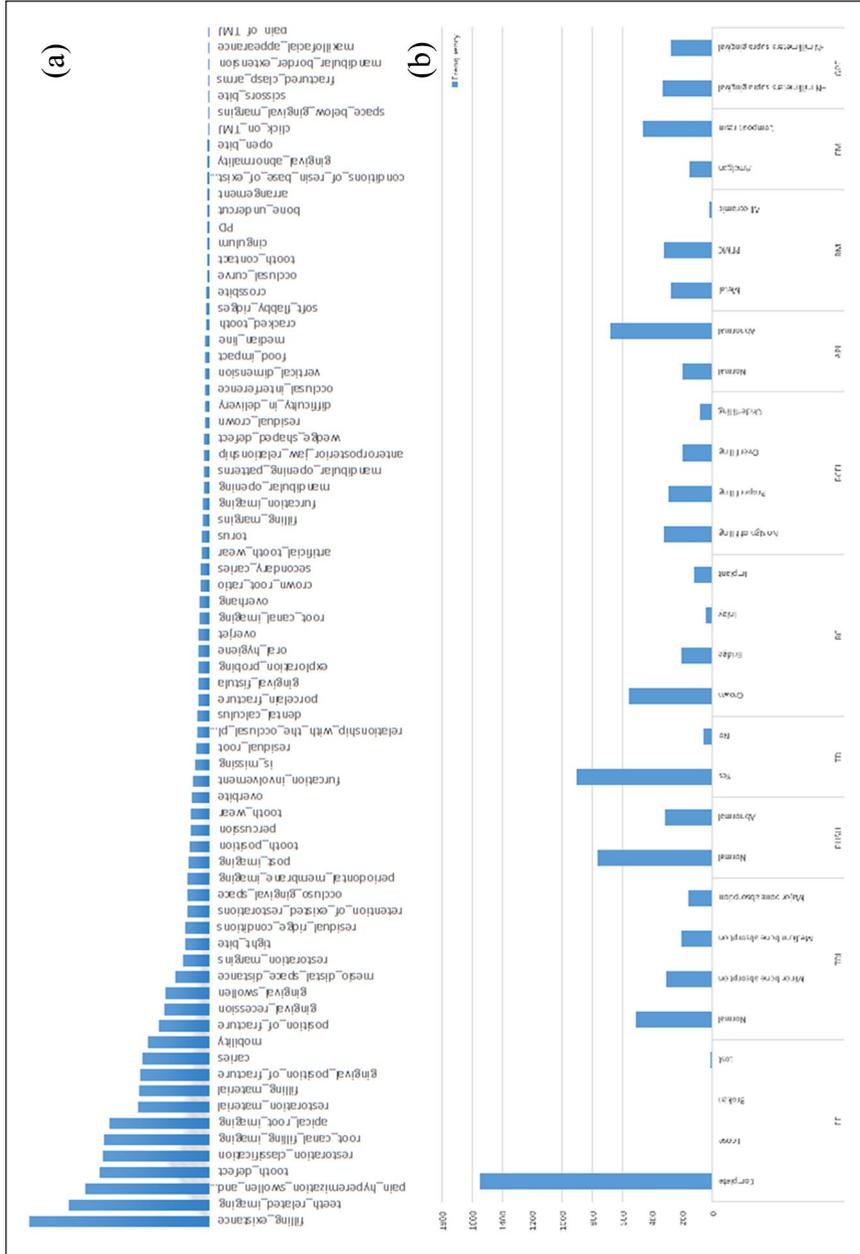$$P = \frac{tp}{tp + fp}, \tag{2}$$

Figure 4. (a) Distribution of attribute frequencies. Obvious differences among attribute frequencies suggest the effect of different attributes on performance of our model, and (b) distribution of attribute value frequencies. Attribute value frequencies of the top 10 attributes show value frequency distributions in each attribute.

$$R = \frac{tp}{tp + fn}, \tag{3}$$

$$F = \frac{2PR}{P + R}, \tag{4}$$

with tp being true positive, fn false negative, and fp false positive.

The tp indicates the number of extracted items (attributes or values) that are identical to annotated items, fn indicates the number of annotated items that remain unextracted, and fp refers to the number of extracted items that remain unannotated.

Moreover, we defined two extra metrics, P_{all} and R_{all}, to reflect overall precision and recall by incorporating frequencies of attributes and attribute values.

Equations are as follows,

$$P_{all} = \sum_{i \in J} P_i * f_i / \sum_{i \in J} f_i, \tag{5}$$

$$R_{all} = \sum_{i \in J} R_i * f_i / \sum_{\in J} f_i, \tag{6}$$

where J is the set of all attributes, i refers to the i_{th} attribute in set J, and f_i refers to the frequency of the i_{th} attribute in set J.

We compared the performance of keyword-based, deep learning and the hybrid methods as illustrated in Tables 1 and 2. We reviewed 50 most frequent attributes and associated values from the overall 88 attributes to calculate P_{all} and R_{all} (Figure 4(a) and (b)). Performance of attribute values of a given attribute are evaluated as a whole.

We listed *P, R*, and *F* score in three methods of the 10 most frequent attributes in Table 1. Sentence2vec performs better than BERT in 9 of the 10 attributes and achieves higher P_{all} and R_{all}. In general, the hybrid method outperforms the other two methods for its overall *F* score is the highest as 0.83, P_{all} is competitive as 0.94 and R_{all} is 0.74. Table 2 shows the performance of 10 most frequent attribute values. The hybrid method is superior in R_{all} as 0.82 and *F* score as 0.88 compared to the keyword-based method.

Discussion

Information extraction aims to extract problem-specific information and then convert it into structured form which can be used directly by classifiers.[19] Due to several disadvantages of clinical texts such as ungrammaticality, abounding with shorthand, being misspelling and unstructured, it poses greater challenges to information extraction on free-text.[1]

Comparison with prior work

Some methods have been developed to address narrative text problems, but challenges still remain.[20] Most medical information extraction tasks require recognition of part of the context, including key phrases and words. In our study, 88 attributes are required to be extracted which almost covers all clinical text in the OES and dramatically increases difficulty and complexity. The keyword-based method is classically used in information extraction and characterized by its high

Table 1. Performance of three approaches for attribute recognition.

Attribute	Description of attributes	P_k	Precision			R_k	Recall			F_k	F1 score		
			P_s	P_b	P_h		R_s	R_b	R_h		F_s	F_b	F_h
Filling existence	The completeness of filling material on a tooth. Whether it is complete, loose, broken or lost.	1	0.87	0.74	0.99	0.32	0.88	0.35	0.57	0.48	0.87	0.48	0.72
Tooth-related imaging	The extent of alveolar ridge resorption of a tooth shown on X-rays, indicating the periodontal health status of the tooth.	0.99	0.96	0.95	0.99	0.94	0.96	0.76	0.97	0.97	0.96	0.85	0.98
Pain hyperemization swollen and ulcer	To measure whether mucosa is on healthy conditions	0.92	0.73	0.66	0.92	0.17	0.83	0.81	0.5	0.28	0.78	0.72	0.65
Tooth defect	To measure whether the tooth is complete or defect in appearance	1	0.85	0.69	0.96	0.37	0.68	0.56	0.61	0.54	0.76	0.63	0.74
Restoration classification	The type of the fixed prosthesis on a tooth	1	0.76	0.63	0.92	0.94	0.84	0.61	0.91	0.97	0.80	0.62	0.92
Root canal filling imaging	To evaluate whether the root canal therapy is complete shown on X-rays	0.99	0.76	0.84	0.98	0.86	0.69	0.82	0.86	0.92	0.73	0.83	0.91
Apical root imaging	To evaluate whether there exists shadow around apical roots on X-rays	1	0.91	0.80	0.98	0.07	0.94	0.86	0.57	0.13	0.92	0.83	0.72
Restoration material	Material of fixed prosthesis on a tooth	1	0.78	0.67	0.96	0.65	0.78	0.67	0.69	0.78	0.78	0.67	0.80
Filling material	Material type of fillings on a tooth	0.95	0.71	0.47	0.92	0.95	0.79	0.75	0.88	0.95	0.75	0.58	0.9
Gingival fracture position	To measure the fractured position of a tooth is below, align or above the gingival.	1	0.82	0.53	0.97	0.14	0.97	0.62	0.68	0.25	0.89	0.57	0.80
All		0.95	0.79	0.69	0.94	0.59	0.80	0.74	0.74	0.73	0.79	0.69	0.83

P_k: precision of keyword-based method; P_s: precision of Sentence2vec method; P_b: precision of BERT method; P_h: precision of hybrid method; R_k: recall of keyword-based method; R_s: recall of sentence2vec method; R_b: recall of BERT method; R_h: recall of hybrid method; F_k: F score of keyword-based method; F_s: F score of sentence2vec method; F_b: F score of BERT method; F_h: F score of hybrid method.

Table 2. Performance of two approaches for attribute value recognition.

Attribute value	Attribute	Precision		Recall		F score	
		P_k	P_h	R_k	R_h	F_k	F_h
Complete/loose/Broken/lost	Filling existence	0.95	0.96	0.87	0.96	0.91	0.96
Normal/minor bone resorption/ moderate bone resorption/ major bone resorption	Teeth related imaging	0.84	0.84	0.62	0.75	0.71	0.79
Normal/abnormal	Pain hyperemization swollenand ulcer	1.00	1.00	0.91	0.98	0.95	0.99
Yes/no	Tooth defect	0.95	0.95	0.98	0.99	0.96	0.97
Crown/bridge/inlay/implant	Restoration classification	1.00	1.00	0.51	0.56	0.67	0.72
No sign of filling/proper filling/ over filling/underfilling	Root canal filling imaging	0.90	0.89	0.72	0.73	0.80	0.81
Normal/abnormal	Apical root imaging	1.00	1.00	0.76	0.78	0.86	0.88
Metal/PFMC/all ceramic	Restoration material	0.90	0.90	0.89	0.97	0.89	0.93
Amalgam/composite resin	Filling material	1.00	1.00	0.54	0.98	0.70	0.99
+N millimeters supra gingival/ −N millimeters subgingival	Gingival position of fracture	1.00	1.00	0.57	0.57	0.72	0.73
All		0.94	0.94	0.76	0.82	0.84	0.88

P_k: precision of keyword-based method; P_h: precision of hybrid method; R_k: recall of keyword-based method; R_h: recall of hybrid method; F_k: F score of keyword-based method; F_h: F score of hybrid method.

precision.[21] When dealing with medical data, this method exhibits a low recall due to the constraints of key phrases and concepts. It also requires significant manual work to conclude keywords, where extensive information must be extracted. Thus, we applied a hybrid method, combining keyword-based and deep learning methods, to extract information, significantly improving the performance relative to the keyword-based method.

Principal results

Fixed-length vectors are learned through neural networks using deep learning models. Mikolov proposed word2vec model to represent sentences,[14] a semantic method that implies sentences with similar meanings comprise vectors with close proximity in multi-dimensional space. Word2vec transforms linguistic words into vectors for computing; Sentence2vec, constructed from Word2vec, learns sentence vectors for each sentence.[13] We ran the models with discretized dimensional vectors (50, 100, and 200 dimensions) separately and found that the models perform best with 100-dimensional vectors representing sentences and 200-dimesional vectors representing words. To the best of our knowledge, we are the first to apply a deep learning method to improve recall of information extraction using Chinese dental clinical data; we have shown its efficiency in recall improvement, which arises from similarity comparison that could distinguish words in similar semantics. Vectors allow multiple expressions of the same attribute to be recognized through similarity calculation. Notably, the attribute *Pain hyperemization swollen and ulcer* is labeled with two keywords, achieving 0.17 of recall; combining with sentence vector, recall improves to 0.5.

BERT is also a neural network-based model for language representation which can be pre-trained from unlabeled text by jointly conditioning on both left and right context in all layers.[12] We tried the BERT model to represent sentences which performs inferior to Sentence2vec (Table 1). The reason is that we used the open source Chinese BERT pre-training model based on corpus in general domain, which brings domain mismatch problem and results in poor performance compared to Sentence2vec-based vectors. The training of BERT model requires much more data than Sentence2vec model and our dental corpus cannot fulfill the needs to train a mature BERT model due to the limited data quantity.

Descriptions of attribute values are clear and simple, suitable for the keyword-matching method. Following word segmentation, vectors of all words are acquired and words with high similarity in vectors are added into the keyword database. An expanded keyword database yields higher recall in attribute value recognition.

In order to investigate a proper training set, we varied the number of EHRs and calculated the associated precision and recall, as shown in Figure 3(a). The increasing number of training sets supports the vector performance of our model. With an increasing number of training sets, overall model performance improves. We assume that the larger training set contains more descriptions of attribute values, thereby improving the recall of our model. Thus, collecting additional training samples will improve our study. Therefore, we set the training set as 1600 annotated EDRs. Evaluation reveals that the hybrid method outperforms the keyword-based and deep learning methods. In attribute recognition (Table 1), P_{all} is consistent between keyword-based and hybrid method. is greatly improved in the hybrid method, relative to the keyword-based method, indicating increased efficiency in the hybrid method. In attribute value extraction (Table 2), recall improves after keyword database expansion, particularly when keyword number increases from 366 to 473. Tables 1 and 2 reveal that recalls of some attributes and values are below R_{all}. For example, attributes Filling Existence (FE), Pain Hyperemization (PH), Gingival Fracture Position (GFP), and Apical Root Imaging (ARI) exhibit recalls of 0.57, 0.5, 0.68, and 0.57. Clearly, the training set is not large enough, resulting in a limited number of sentences and words in some attributes and values. It is common to tradeoff between recall and precision. It is true that a maximum is obtained with 0.8 threshold value considering the F1 score. However, we tend to consider the precision is more important as mentioned in our paper. The reason is that a false attribute or value could have a worse impact on the CDSS output. And the CDSS would pretend missed attributes as attributes with default values. Therefore, we chose a higher threshold as 0.9. The high similarity thresh old achieves high precision and low recall results. In Table 1, the recalls of FE, PH, GFP, and ARI increase dramatically in the deep learning method, relative to the hybrid method. Alternative approaches might be implemented to increase recall, including expansion of the training set and adjustment of the threshold for different attributes.

Error analysis

Reviewing all error predict results, we found there exist some patterns of mistakes that were summarized below in detail. Cases of different errors are exemplified in Table 3.

   (1) Sentences match more than one attribute.

The sentences that are labeled with more than one attribute always are recognized with one attribute. The reason is that in our framework, we regard each sentence as a binary problem, and only

Table 3. Examples of four error types.

Error type	Example in English	Predict data property	Correct data property	Notes
Type 1	Tooth 17 (FDI notation) has white filling on the mesial occlusal side.	Filling material	Filling material; filling on distal surface	
Type 2	Percussion (–)	Filling existence	Percussion	Previously sentence: there exist white filling on the mesial occlusal side.
Type 3	The fracture of tooth is below gingiva 2–3 mm	Position of fracture	Gingival position of fracture	
Type 4	Metal crown on 37; margin fitness is good.	Restoration material	Restoration material; restoration margin	The segmentation tool did not divide the text into two sentences.

extract one attribute with the highest similarity. For example, sentence "Tooth 17 (FDI notation) has white filling on the mesial occlusal side" labeled with two attributes is matched to one of them in our study.

(2) Sentences tend to be recognized with attributes of sentences in neighborhood.

This phenomenon is in relevant with the training model of vectors. In the unsupervised models, vectors are trained to take sentences nearby into consideration. Therefore, there exist some confusions to some extent.

(3) Sentences where the incorrect attributes matched were, in semantic, highly similar to sentences with these properties.

This is probably due to the nature of the vector training algorithms and keywords embedded. Sentences with high resemblance tend to be calculated high cosine similarities and recognized with wrong properties.

(4) Segmentation problem

A few sentences are segmented incorrectly with extra symbols, like (–), /. This has an impact on vectors training and results in wrong extraction.

Conclusion

We proposed a novel NLP workflow combining keyword-based and deep learning methods in extracting attributes and values from EDRs. Evaluation results indicate that the hybrid method outperforms both the keyword-based method and deep learning method. The workflow could be potentially utilized as an initial step to provide structured data for data-based models training. In future we will integrate our NLP workflow with data-driven CDSSs and apply it to more corpuses from diverse sources for more general use.

Author contributions

Declaration of conflicting interests

Funding

ORCID iD

Qingxiao Chen iD https://orcid.org/0000-0002-7727-251X

References

1. Meystre SM, Savova GK, Kipper-Schuler KC, et al. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008; 47: 128–144.
2. Hypponen H, Saranto K, Vuokko R, et al. Impacts of structuring the electronic health record: a systematic review protocol and results of previous reviews. *Int J Med Inform* 2014; 83: 159–169.
3. Demner-Fushman D, Chapman WW and McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009; 42: 760–772.
4. Nadkarni PM, Ohno-Machado L and Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011; 18: 544–551.
5. Chen Q, Wu J, Li S, et al. An ontology-driven, case-based clinical decision support model for removable partial denture design. *Sci Rep* 2016; 6: 27855.
6. Schleyer TK, Ruttenberg A, Duncan W, et al. An ontology-based method for secondary use of electronic dental record data. *AMIA Summits Transl Sci Proc* 2013; 2013: 234–238. http://europepmc.org/abstract/MED/24303273; https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/24303273/?tool=EBI; https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/24303273/pdf/?tool=EBI; https://europepmc.org/articles/PMC3845770; https://europepmc.org/articles/PMC3845770?pdf=render (2013, accessed 18 March 2013).
7. Chapman WW, Nadkarni PM, Hirschman L, et al. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011; 18: 540–543.
8. Liu K, Hogan WR and Crowley RS. Natural language processing methods and systems for biomedical ontology learning. *J Biomed Inform* 2011; 44: 163–179.
9. Christensen LM, Harkema H, Haug PJ, et al. ONYX: a system for the semantic analysis of clinical text. In: The workshop on current trends in biomedical natural language processing, Boulder, CO: Association for Computational Linguistics, 2010, pp.19–27.
10. Irwin JY, Harkema H, Christensen LM, et al. Methodology to develop and evaluate a semantic representation for NLP. *AMIA Annu Symp Proc* 2009; 2009: 271–275.
11. Che W, Li Z and Liu T. LTP: a Chinese language technology platform. In: *Proceedings of the 23rd international conference on computational linguistics: demonstrations*, Beijing, 2010, pp.13–16. Association for Computational Linguistics.

12. Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv: Comput Lang*, 2018.
13. Le QV and Mikolov T. Distributed representations of sentences and documents. *Proc Mach Learn Res* 2014; 32(2): 1188–1196.
14. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. *Adv Neural Inform Process Syst* 2013; 26: 3111–3119.
15. Mikolov T, Le QV and Sutskever I. Exploiting similarities among languages for machine translation. *Comp Sci* 2013. http://arxiv.org/abs/1309.4168
16. Artstein R. Inter-annotator agreement. In: Ide N and Pustejovsky J (eds.) *Handbook of linguistic annotation*. Dordrecht: Springer, 2017, pp.297–313.
17. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. *Comp Sci* 2013. https://arxiv.org/abs/1301.3781v3
18. Spooner SA. Mathematical foundations of decision support systems. In: Berner ES (ed.) *Clinical decision support systems: theory and practice*. New York, NY: Springer, 2007, pp.23–43.
19. Nadkarni PM, Ohno-Machado L and Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011; 18: 544–551.
20. Zhang S, Kang T, Zhang X, et al. Speculation detection for Chinese clinical notes: impacts of word segmentation and embedding models. *J Biomed Inform* 2016; 60: 334–341.
21. Kang N, Singh B, Afzal Z, et al. Using rule-based natural language processing to improve disease normalization in biomedical text. *J Am Med Inform Assoc: JAMIA* 2013; 20: 876–881.

Appendix

Unlike ontologies previously built,[6] we designed an ontological paradigm to represent concepts and terms in the CDSS, which produces RPD designs by analyzing patients' oral conditions. The ontology we built describes specific oral conditions of partial edentulism and RPD design treatments through defining 70 classes, 203 data properties, and 48 object properties.

Among those classes and properties, one major content is about patients' oral conditions of each visit. Data properties among the ontology represent oral conditions, values of which represent results of the corresponding oral examinations. Forty-eight object properties describe interrelations among oral examinations, tooth positions, and RPD components. The study here utilizes classes and properties regarding oral conditions to extract information from the OES of EDR data.

In the ontology, classes represent upper levels of oral examinations; their data property represent lower levels of oral examinations. The lowest level of the data properties represent the detailed oral conditions, which directly link to the content of the OES in EDRs. Here we will introduce classes and properties related to this study in brief.

Define three classes representing top levels of oral examinations: *oral conditions* (representing the top level of oral examinations), *tooth* (representing oral examinations related to tooth positions), *mouth* (representing oral examinations of the whole mouth).

Define levels of data properties representing specific oral examinations in hierarchy. Table 4 shows parts of data properties of *Class tooth*. Only the lowest levels of data properties were defined data property values.

Define object properties representing relations among classes and data properties, namely *has_part, is_part_of, tooth_object_property, left_first_tooth* and *right_first_tooth*. Figure 5 shows an instance of the semantic network of oral conditions of partial edentulism.

Table 4. Exemplary data properties of *Class tooth*.

|	Data property
Level 1	General conditions (define general conditions)
Level 1-2	Is missing
	Tooth defect
	Residual crown
	Residual root
Level 2	Restorative and endodontic examination
Level 2-1	Endodontic examination
Level 2-1-1	Tooth food impact
	Tooth wear
	Percussion
	Exploration probing
	Position of fracture
	Gingival position of fracture
	Caries
Level 2-2	Filling examination
Level 2-2-1	Filling existence
	Filling material
	Filling margins
Level 2-3	Restoration examination
Level 2-3-1	Restoration classification
	Restoration material
Level 2-4	Periodontal examination
Level 2-4-1	Mobility
	PD
	Gingival recession
	Dental calculus
	Gingival swollen
Level 2-5	Location of residual teeth
Level 2-5-1	Tooth position
	Relationship with the occlusal plane
Level 2-6	Imaging
Level 2-6-1	Teeth related imaging
	Root canal filling imaging
	Apical root imaging
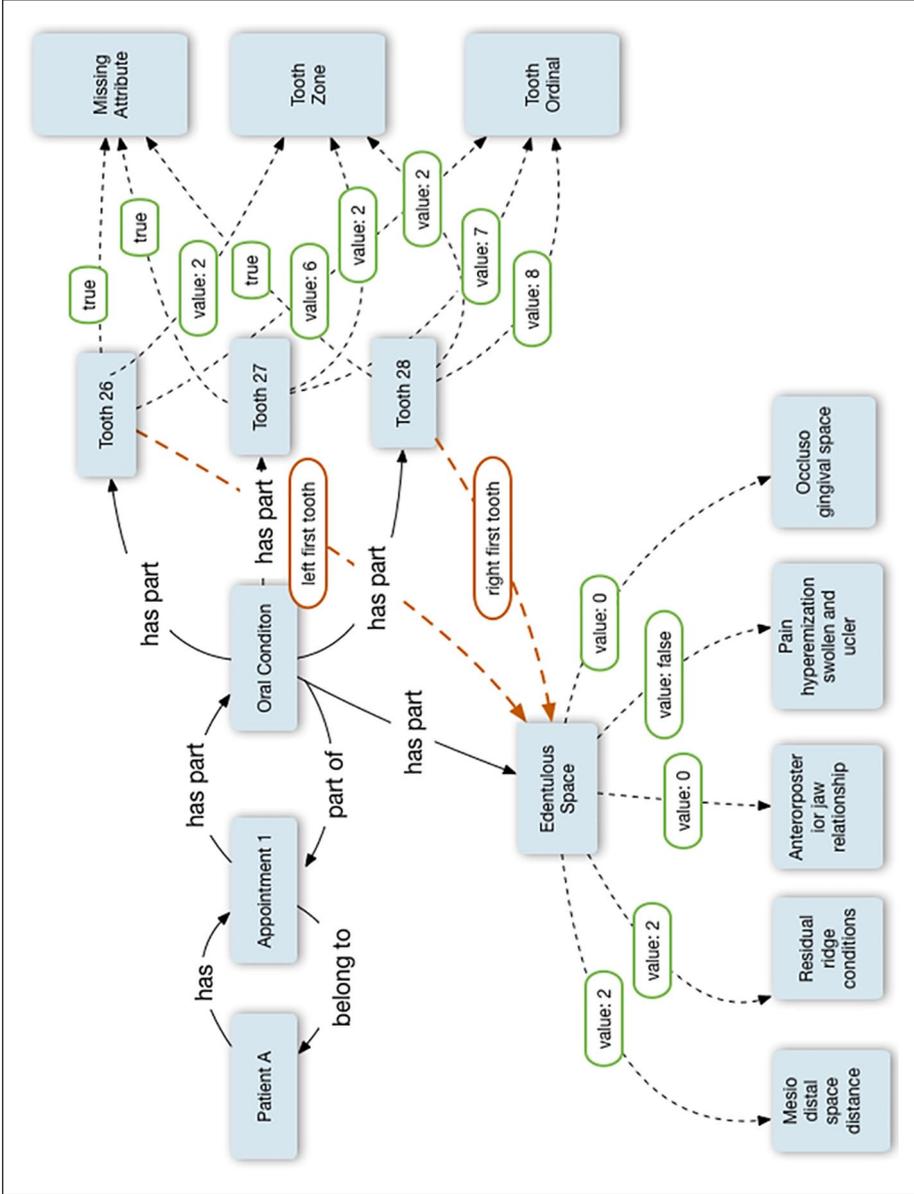
Figure 5. Network of an instance in ontology. An instance of Patient A was established, in which classes and attributes were connected through interrelations.

Annotation guidelines

As the ontology defines oral examinations and the logic relations, annotators only need to map phrases and symbols to the terminal node of the semantic network representing the lowest levels of data properties and their values during the annotation process. Most sentences comprise of tooth positions, examinations and their results; thus, each sentence was annotated with data properties and values respectively. For example, in the annotated sentence "<pos>6/8765/4567/18剩余牙槽嵴中度吸收, {conditions of residual ridge: 2; pos:$nt}
" (6/8765/4567/18 refers to teeth 16, 24, 25, 26, 27, 31, 38, 45, 46, 47, and 48 in FDI notation; the Chinese sentence refers to moderate resorption of residual ridge), "<pos>" marks tooth positions, and text in the curly brace represents data properties and their values in ontology. Table 5 shows one record of an anonymous patient's visit translated into English language from Chinese language. Table 6 shows the annotation data of the example in Table 5.

Table 5. One electronic health record from the Prosthodontics database, translated into English.

Chief complaint	A tooth on the upper right jaw has been defective for >3 years.
Medical of presenting illness	One tooth on the upper right jaw underwent root canal therapy 3 years ago. The patient claims part of the filling material was missing 1 week ago and requires restoration treatment.
Examination of oral conditions	Tooth 17 (FDI notation) has white filling on the mesial occlusal side. Part of the filling is missing. Percussion (–). Mobility 1°. There exists swallow on the gingival of tooth 17. X-ray picture shows proper root filling and there is no abnormal sign around the apical roots. Teeth 14, 15, 24, 36, 37, 38, 45, 46, and 47 are missing. Residual ridge suffers moderate bone resorption. Mesial-distal distance is sufficient. Vertical distance is sufficient.
Diagnosis	Tooth defect of 17; partial edentulism on maxilla and mandible.
Treatment plan	Crown for tooth 17; removable partial denture for partial edentulism.

Table 6. Annotations of the OES from the exemplary EDR.

Text in the OES	Tooth position in FDI annotation	Data property (Attribute)	Value	Notes of values
Tooth 17 (FDI notation) has white filling on the mesial occlusal side.	17	Filling material	1	White (composite resin/glass ionomer)
Tooth 17 (FDI notation) has white filling on the mesial occlusal side.	17	Filling on distal surface	1	True
Part of the filling is missing.	17	Filling existence	1	Damaged or partially missing
Percussion (–)	17	Percussion	0	Negative symptom
Mobility 1°	17	Mobility	1	One degree
There exists swallow on the gingival of tooth 17.	17	Gingival swollen	1	True
Teeth 14, 15, 24, 36, 37, 38, 45, 46, and 47 are missing.	14,15,24,36,37, 38,45, 46, 47	Missing	1	True
Residual ridge suffers moderate bone resorption.	14, 15, 24, 36, 37, 38, 45, 46, 47	Residual ridge conditions	2	Medium degree
Mesial-distal distance is sufficient.	14, 15, 24, 36, 37, 38, 45, 46, 47	Mesio-distal space distance	0	moderate
Vertical distance is sufficient.	14, 15, 24, 36, 37, 38, 45, 46, 47	Occlusal gingival space	0	moderate