The Model for Pneumothorax Knowledge Extraction Based on Dependency Syntactic Analysis

Xiangge Liu[1], Jing Li[1](✉), and Yuzhuo Zhao[2]

[1] School of Economics and Management, Beijing Jiaotong University, Beijing, China
jingli@bjtu.edu.cn

[2] Department of Emergency, Chinese PLA General Hospital, Beijing, China

Abstract. As the frontier application technology of artificial intelligence in the medical field, medical knowledge mapping plays an important role in assisting medical staff in the diagnosis and treatment of diseases. In this paper, pneumothorax is selected as the research object, and the method of natural language processing is applied to study the knowledge extraction model involved in the construction of knowledge map. Firstly, based on the principle of dependency syntax, the grammatical relationship of pneumothorax corpus is analyzed, and the dependency syntax tree is constructed. Secondly, based on the special text features of pneumothorax corpus, a knowledge extraction model based on Text Syntactic Structure and dependency representation is proposed, and the predicate is used as the core term to extract pneumothorax triples. Finally, based on the standardized knowledge representation defined by experts, the automatically extracted triples are normalized and presented in the form of pneumothorax knowledge map.

Keywords: Knowledge graph · Knowledge extraction · Dependency syntactic analysis · Pneumothorax

1 Introduction

With the development of artificial intelligence, the combination of technologies such as natural language processing and knowledge mapping can help clinicians and patients by extracting knowledge from medical data such as medical textbooks, medical encyclopedias and clinical cases. The main purpose of knowledge extraction is to extract structured knowledge from raw semi-structured or unstructured text through a series of methods or means, and express it as a triple process of “entity-relationship-entity” [1]. To a certain extent, the effectiveness of knowledge extraction determines the accuracy of knowledge application.

In the aspect of knowledge extraction in the medical field, Seol [2] recognized the entities related to clinical manifestations such as symptoms and examination results by using sequence annotation model, and extracted the interaction and relationship between diseases by using support vector machine model. Savova [3] extracted the clinical manifestations of cancer from clinical texts by natural language processing. Wei

X. Shi et al. (Eds.): *LISS 2021*, LNOR, pp. 160–168, 2022.
https://doi.org/10.1007/978-981-16-8656-6_15

[4] used natural language processing technology to extract the names of chemical drugs and diseases in biomedical articles, and completed the task of automatically extracting whether chemical drugs have pathogenic relationship with diseases. At present, knowledge extraction in the medical field mainly includes the extraction of specific targets, such as disease name and symptoms [3], or knowledge extraction for a certain kind of disease to supplement the knowledge base or build knowledge map [5]. There are often some shortcomings, such as the scope of knowledge extraction is too limited to meet the medical needs of specific scenarios.

This article takes the emergency medical protection of the Winter Olympic Games as the target scene, selects the emergency injury condition of pneumothorax as the research object, extracts relevant data from relevant medical books under the guidance of professional doctors, and applies natural language processing methods to extract knowledge of pneumothorax. Based on this, the knowledge map of pneumothorax is constructed, so as to provide auxiliary decision support for the diagnosis and treatment of pneumothorax and meet the specific medical needs in actual scenarios.

2 Knowledge Extraction Process and Model Construction

2.1 Knowledge Extraction Process Based on Dependency Syntactic Analysis

Syntactic analysis is one of the key techniques in natural language processing. As there are syntactic structures in linguistic expressions, such as subject-verb-object, verb-object and subject-subject structures, syntactic analysis supports natural language processing tasks by analyzing the semantic structure of texts in a natural language understanding manner [6]. Dependent syntax states that the predicate in a sentence is generally a statement of the subject, indicating “what”, “what” and “how”, and plays a key role in grammatical expressions.

In this paper, based on the principle of dependent syntactic analysis, combined with the textual features of the original pneumothorax data, the pneumothorax knowledge triad is extracted by formulating knowledge extraction rules, as shown in Fig. 1 for the pneumothorax knowledge extraction process based on dependent syntactic analysis, the main ideas are as follows.

(1) Analysis of the dependency syntactic relations between the texts of the original pneumothorax data.
(2) Based on the results of the dependency analysis of the pneumothorax corpus, a dependency syntactic tree is constructed for the dependent lexical items and specific relations.
(3) Analyze the special syntactic features of the pneumothorax text and use them as the basis for formulating knowledge extraction rules.
(4) According to the extraction rules, the entities and their relations with associated relations in the pneumothorax knowledge are extracted in a triad.
(5) Based on the requirements of knowledge map storage and application, the triad of pneumothorax knowledge is formatted and normalized, and presented in the form of pneumothorax knowledge map.

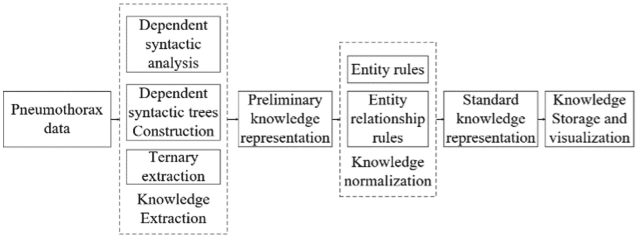

Fig. 1. Pneumothorax knowledge extraction process

2.2 The Knowledge Extraction Model Based on Text Features and Dependencies

Based on the syntactic analysis of dependency, this paper proposes a knowledge extraction method based on the syntactic structure of the original pneumothorax text combined with dependency relations, and the extraction rules are shown below as the core of knowledge extraction.

Rule 1: For subject-predicate object relations. The subject-verb-object triad of "subject-predicate-object" is extracted directly from the predicate as the core of the description of the relationship between subject-predicate (SBV) and verb-object (VOB) on the left and right sides respectively.

Rule 2: For object-first subject-predicate relations. First identify the prepositional object (FOB) in the text, find the central word in the object relationship, and then use the predicate as the core in combination with the verb-object relationship for the extraction of the triad.

Rule 3: For the definite postpositional verb-object relationship. First identify the definite middle relation (ATT) in the text, find the central word in the definite middle relation, and then use the predicate as the core for the extraction of the triad according to Rule I.

Rule 4: For subject-verb-movement-complement relations that contain a prepositional relationship. The head entity and the relation are first identified by the subject-predicate relation (SBV) and the verb-complement structure (CMP), and then the tail entity is identified by the prepositional-object relation (POB).

Rule 5: For the treatment of juxtaposition structures. Identify juxtapositions (COO) in the text and perform triadic extraction according to the principle that juxtaposed syntactic relations share syntactic components, combining the four rules above.

3 Pneumothorax Knowledge Extraction Based on a Dependent Feature Extraction Model

3.1 Data Sources and Pre-processing

Based on the professionalism of medical knowledge, several professional medical books on pneumothorax related knowledge in Internal Medicine [7], Huang Jiasi Surgical [8] and Emergency Manual [9] were selected as the main data sources under the advice of professional doctors, and supplemented by pneumothorax knowledge in the third-party medical website Seeking Medical Advice (https://www.xywy.com) to obtain pneumothorax including The original data on pneumothorax including disease introduction, etiology, symptoms, treatment, etc. were obtained.

Word segmentation and lexical tagging are the basis of knowledge extraction, which is to segment Chinese text into a single word or phrase through a specific method, and then annotate it, which is of great significance to text information analysis. The module package of word segmentation and lexical tagging provided by LTP uses customized pneumothorax proper noun dictionary in word segmentation stage and 863 lexical tagging set in lexical tagging stage to improve the separation efficiency of pneumothorax knowledge. In the lexical annotation stage, 863 lexical annotation set is used to annotate the original pneumothorax text.

3.2 Syntactic Analysis of Dependencies

This paper applies the principles of dependent syntactic analysis, based on the grammatical-logical structure of the sentence and based on LTP, on the basis of participle and lexical annotation.

The LTP dependency syntactic analysis model provides 14 types of dependency syntactic relationship structures, including subject-verb, verb-object, and definite-medium relationships, etc. This paper analyses the semantic relationships of the lexical items in the original pneumothorax text based on LTP. As shown in Fig. 2, the process of dependency syntactic analysis of pneumothorax text is demonstrated.

For the given sentence "Pneumothorax's typical symptoms is sudden chest pain, followed by breathlessness." The text is analyzed for dependencies between lexical items, where the predicate "is" and "followed" are identified as the core item (Root), which is the key to triadic extraction, and the lexical and dependency characteristics of each item are used to determine the pairs of items that are important to the expression of the text.

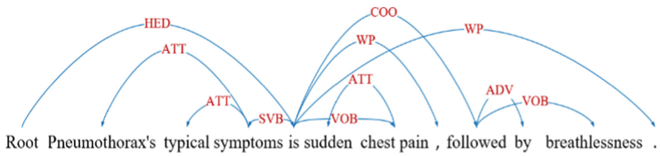

Fig. 2. Example of dependent syntactic analysis

Based on the need for a triadic extraction task, a dependency syntactic tree is constructed for the lexical items with dependencies and specific relations based on the results of the dependency analysis of the pneumothorax text. The main process is as follows, for the pneumothorax raw text collection T = $\{t_1, t_2, \ldots\ldots, t_n\}$, t_i, t_j are dependent lexical items, which represented by the dependency pair $\left(t_i, r, t_j\right)$, where r is the dependency arc from t_i to t_j. For any dependency pair $\left(t_i, r, t_j\right)$, the dependency parent node t_i and the dependency relation r are extracted and a lexicon is built, and when traversing the dependency syntax tree, the text dependency pair is matched by indexing the dependency parent node.

3.3 Knowledge Triad Extraction

The set of dependency pairs of lexical items with dependency relations in the text can be obtained by traversing the dependency syntax tree. However, due to the characteristics of the original text description of the pneumothorax, the dependency pairs triad obtained through the dependency syntax analysis cannot be used directly. In the process of knowledge triad extraction, a knowledge extraction model based on text features and dependency relations is combined to extract the triads of entities and entity relations with associated relations in pneumothorax knowledge according to the knowledge extraction rules.

In the process of pneumothorax knowledge extraction, a combination of multiple rules is often used for extraction, and the result obtained is a preliminary knowledge representation in the form of "entity-relationship-entity", where the head entity and tail entity represent the lexical items with dependencies, and the relationship is the dependency between them. The results of pneumothorax knowledge extraction are shown in Table 1.

Table 1. Example of pneumothorax knowledge extraction results

Original text	Knowledge extraction results
Pneumothorax's typical symptoms is sudden chest pain,followed by breathlessness	Pneumothorax's typical symptoms–is–sudden chest pain
	Pneumothorax's typical symptoms–followed–breathlessness
Pneumothorax patients are prohibited from travelling by air	Pneumothorax patients–prohibited–travelling by air
Pneumothorax can be divided into three categories: spontaneous pneumothorax, traumatic pneumothorax and pneumothorax of medical origin	Pneumothorax–divided–spontaneous pneumothorax
	Pneumothorax–divided–traumatic pneumothorax
	Pneumothorax–divided–pneumothorax of medical origin

3.4 Standardized Knowledge Representation

The triad of pneumothorax knowledge automatically extracted based on dependent syntactic analysis is deficient in terms of the normality of the presentation and cannot be directly utilized. Firstly, there is some variability in the description of pneumothorax-related knowledge due to the two different data sources in this paper. Secondly, as medical texts have unique representations, the form of automatically extracted knowledge is not very standard. For example, for the relationship "pneumothorax"-"symptom", as the predicates are expressed in different ways, it will both For example, for the relationship "pneumothorax" - "symptom", the different ways of expressing the predicates can bring unnecessary redundancy and increase the complexity of understanding. Thirdly, the automatically extracted pneumothorax knowledge in the form of a triad does not have a strong cohesiveness for the presentation of the knowledge graph.

Therefore, this paper defines a set of standardized knowledge representation rules in the form of communication with experts to disambiguate and unify pneumothorax entity concepts, entity types and entity attributes from different sources under the same specification. Some entities and entity types are shown in Table 2.

Table 2. Example of entities and entity types

Entity type	Entity name
Name of injury	Pneumothorax
ICD-10	J93.901
Definition	Pneumothorax
Name of secondary injury	Spontaneous pneumothorax
Name of grade III injury	Closed pneumothorax
Hospital	Beijing Haidian Hospital
Department	Thoracic Surgery
Susceptible person	Male
Susceptible age	20–40 year
Risk factor	Chest injury
Predilection site	Pleural cavity
Symptom	Dyspnea
Sign	Blood pressure drops
Differential diagnosis	Asthma
Complication	Empyema
Laboratory examination	Pulmonary function test
Imaging examination	X-ray examination
Physical examination	Measurement of intrathoracic pressure
Other auxiliary examinations	Thoracoscopy
Medical history	Bullae of lung
First aid measures	Debridement
Basic treatment	Get enough rest

(*continued*)

Table 2. (*continued*)

Entity type	Entity name
Medication	Antibiotic
Surgical treatment	Thoracoscope
Other therapies	Exhaust therapy
Prognosis	Review
Prevention	Avoid smoking

According to the storage requirements of attribute graph model in graph database, in the process of knowledge mapping representation, the node is the corresponding medical entity, the label is the type of entity, and the relationship is the description of relationship between entities. The standard knowledge representation is shown in Table 3.

Table 3. Example of standard knowledge representation

Source node	Source node type	Target node	Target node type	Relationship
Spontaneous pneumothorax	Name of secondary injury	Pneumothorax	Name of injury	Belong
Pneumothorax	Name of injury	Thoracic Surgery	Department	Consultation
Thoracic Surgery	Department	Haidian Hospital of Peking University	Hospital	Recommend
Male	Susceptible person	Pneumothorax	Name of injury	Epidemiology
Chest	Position	Pneumothorax	Name of injury	Position
Smoking	Risk factor	Pneumothorax	Name of injury	Cause
Chest pain	Symptom	Pneumothorax	Name of injury	Symptom
Blood pressure drops	Sign	Pneumothorax	Name of injury	Sign
Pneumothorax	Name of injury	Asthma	Differential diagnosis	Identify
Pneumothorax	Name of injury	Hemopneumothorax	complication	Initiation
Pneumothorax	Name of injury	Bullae of lung	medical history	Medical history

(*continued*)

Table 3. (*continued*)

Source node	Source node type	Target node	Target node type	Relationship
Pneumothorax	Name of injury	X-ray examination	Imaging examination	Inspect
Pneumothorax	Name of injury	Thoracotomy	Surgical treatment	Treatment
Thoracoscope	Surgical treatment	Surgical treatment	Treatment measures	Belong
Pneumothorax	Name of injury	Antibiotic	Medication	Medication

3.5 Pneumothorax Knowledge Storage and Visualisation

Through knowledge extraction, the scattered and unstructured pneumothorax data are aggregated into structured knowledge and presented in a visualization effect, which can clearly and intuitively derive the characteristics related to pneumothorax disease. At the same time, the results of the knowledge extraction are further applied to the pneumothorax knowledge graph and intelligent question and answer system to provide a basis for pneumothorax prevention as well as clinical judgement and treatment. The pneumothorax knowledge is integrated according to the rules of knowledge representation and graph database storage to obtain a pneumothorax entity database containing 27 entity types and 245 entities, and a pneumothorax relationship database with 14 relationship types and 274 relationships. In this paper, the Neo4j graph database is used to store and

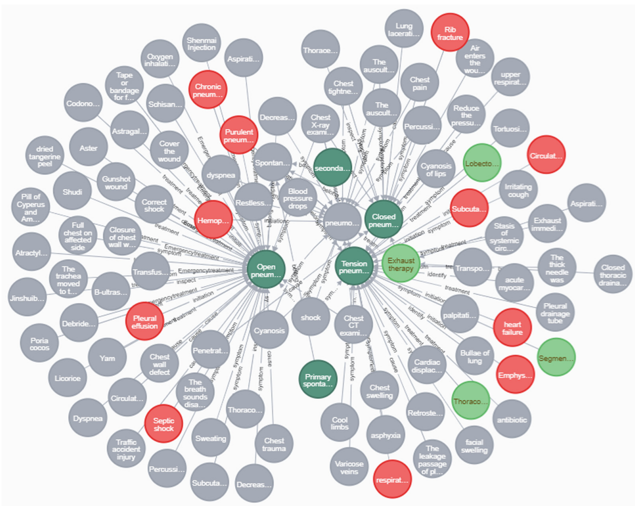

Fig. 3. Example of complete pneumothorax knowledge map

visualize the pneumothorax knowledge, and part of the pneumothorax knowledge graph is shown in Fig. 3.

4 Conclusion

In this paper, pneumothorax is selected as the research object to study the key technology of knowledge extraction in the process of pneumothorax knowledge mapping construction from the perspective of auxiliary diagnosis of injury in emergency medical scene. It is mainly applied to the principle of dependency parsing in natural language processing. By combining the text features of pneumothorax original corpus, a knowledge extraction model based on context syntactic structure and dependency relation is proposed to extract the keyword dependency pairs in pneumothorax text. In order to meet the retrieval requirements of pneumothorax auxiliary diagnosis, according to the representation rules of entity and relation type defined by experts, the automatically extracted triples are represented in the form of standardized knowledge, and presented in the form of knowledge map, which achieves good visualization effect. For future research, structured knowledge extracted from knowledge can be used in pneumothorax knowledge mapping or intelligent question answering system, so as to provide decision support for pneumothorax diagnosis and treatment, and better meet the needs of pneumothorax emergency medical treatment.

Acknowledgment. This work was partly supported by the National Key Research and Development Plan for Science and Technology Winter Olympics of the Ministry of Science and Technology of China (2019YFF0302301).

References

1. Sowa, J.F.: Principles of semantic networks: exploration in the representation of knowledge. Frame Probl. Artif. Intell. **2–3**, 135–157 (1991)
2. Seol, J.W., Yi, W., Choi, J., et al.: Causality patterns and machine learning for the extraction of problem-action relations in discharge summaries. Int. J. Med. Inform. 1 (2016)
3. Savova, G.K., Danciu, I., Alamudun, F., et al.: Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Can. Res. **79**(21), 5463–5470 (2019)
4. Wei, C.H., Peng, Y., Leaman, R., et al.: Overview of the BioCreative V chemical disease relation (CDR) task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop (2015)
5. Yu, T., Li, J., Yu, Q., et al.: Knowledge graph for TCM health preservation: design, construction, and applications. Artif. Intell. Med. **77**, 48–52 (2017)
6. Che, W., Li, Z., Liu, T.: LTP: a Chinese language technology platform. In: COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, Beijing, China, 23–27 August 2010. Association for Computational Linguistics (2010)
7. Ge, J., Xu, J.: [Internal medicine. 8th Edition]. Renmin Weisheng Chubanshe (2013). (in Chinese)
8. Wu, M., Wu, Z.: [Huang Jiasi surgery 7th Edition]. Renmin Weisheng Chubanshe (2008). (in Chinese)
9. Shi, Y.: [Emergency manual - Fourth Edition]. Renmin Weisheng Chubanshe (2002). (in Chinese)