Received: 10 August 2018 | Revised: 20 February 2019 | Accepted: 1 March 2019

DOI: 10.1111/exsy.12401

WILEY Expert Systems

SPECIAL ISSUE PAPER

A knowledge construction methodology to automate case-based learning using clinical documents

Maqbool Ali[1,2] | Jamil Hussain[1] | Sungyoung Lee[1] | Byeong Ho Kang[2] | Kashif Sattar[3]

[1]Department of Computer Science and Engineering, Kyung Hee University, Gyeonggi, 446-701, Yongin, Republic of Korea
[2]School of Engineering and ICT, University of Tasmania, Tasmania, 7005, Hobart, Australia
[3]University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi, Pakistan

Correspondence
Sungyoung Lee, Department of Computer Science and Engineering, Kyung Hee University, Yongin, Gyeonggi, 446-701, Republic of Korea.
Email: sylee@oslab.khu.ac.kr

Funding information
Institute for Information & Communications Technology Promotion (IITP), Ministry of Science and ICT, Republic of Korea, Grant/Award Number: IITP-2017-0-00655 and IITP-2017-0-01629; Ministry of Trade, Industry & Energy (MI), Republic of Korea, Grant/Award Number: 10052955

Abstract

The case-based learning (CBL) approach has gained attention in medical education as an alternative to traditional learning methodology. However, current CBL systems do not facilitate and provide computer-based domain knowledge to medical students for solving real-world clinical cases during CBL practice. To automate CBL, clinical documents are beneficial for constructing domain knowledge. In the literature, most systems and methodologies require a knowledge engineer to construct machine-readable knowledge. Keeping in view these facts, we present a knowledge construction methodology (KCM-CD) to construct domain knowledge ontology (i.e., structured declarative knowledge) from unstructured text in a systematic way using artificial intelligence techniques, with minimum intervention from a knowledge engineer. To utilize the strength of humans and computers, and to realize the KCM-CD methodology, an *interactive case-based learning system* (iCBLS) was developed. Finally, the developed ontological model was evaluated to evaluate the quality of domain knowledge in terms of coherence measure. The results showed that the overall domain model has positive coherence values, indicating that all words in each branch of the domain ontology are correlated with each other and the quality of the developed model is acceptable.

KEYWORDS

case-based learning, clinical case, controlled natural language, declarative knowledge, knowledge engineering, ontological model

1 | INTRODUCTION

Case-based learning (CBL) is an active learning approach that provides a favourable context for students to explore, question, discuss, and share their experiential knowledge to improve their practical intelligence (Demircioğlu & Selçuk, 2016). CBL has maintained its focus around clinical, communal, and scientific problems and is not a new term as it has been used in the medical domain since 1912 (McLean, 2016). In terms of student-centric pedagogy, CBL is being widely used in various health care training environments around the world (Eseonu, Carachi, & Brindley, 2013, Gade & Chari, 2013; Osinubi & Ailoje-Ibru, 2014; Patil & Karadesai, 2016; Sule, 2016). This approach has been met with general acceptance in the fields of medicine, dentistry, pharmacology, occupational and physical therapy, nursing, allied health fields, and child development. Similarly, this approach has been utilized in various departments, including medical education, information technology, and quality improvement, and has been practiced in rural as well as underserved areas (McLean, 2016). Findings have validated the effectiveness and universal nature of CBL, which is especially useful for the curricula of medical and health professions (McLean, 2016).

CBL is a student-centric teaching methodology that makes use of problem-based learning principles. For problem-based learning, humans and computers can play a key role in the medical domain. However, both have strengths and weaknesses (Cummings, 2014; Halim, 2018). (a) Human judgement is considered credible, (b) humans have common sense and can determine new rules, and (c) humans can easily identify trends or abnormalities in visualization data. However, humans also have weaknesses whereby they (a) often cannot accomplish complex computational decisions, (b) cannot perform fast reasoning computations, and (c) get easily tired and bored. These human weaknesses can be mitigated by using a computer, which can perform complex computation decisions relatively faster and will not suffer from tiredness or boredom.

In health care professional education, students tackle uncertain situations due to the interplay of a number of problems (Baillergeau & Duyvendak, 2016). In such situations, each student has his/her own judgement, opinion, and feedback and consider this integral and appropriate for that situation. In such situations, experiential knowledge is considered a resource (Baillergeau & Duyvendak, 2016) that can facilitate and provide lived knowledge to students. According to Willoughby and Philosophy (2018), "Experiential knowledge is a knowledge of particular things gained by perception and experience." Experiential knowledge enables individuals to capture practical experience for problem-solving. It is considered as a valuable resource to enhance an individual's participation and user empowerment (Baillergeau & Duyvendak, 2016).

Students are subject to human weakness to become tired or bored and tend to choose computer-based cases as opposed to lectures for their learning (Gopalan, 2016; Thistlethwaite et al., 2012). Additionally, to support learning outcomes, a large number of web-based learning systems have been developed (Ali, Bilal, Hussain, Lee, & Kang, 2015; Boubouka, 2013; Cheng, Sheng-Huang, Shi-Jer, & Ru-Chu, 2012; Chen, Cheng, Sheng-Feng, Yong-Guo, & Lin, 2009; Shyu, Liang, Hsu, Luh, & Chen, 2004; Suebnukarn & Haddawy, 2007; UTMB, 2013; UNM, 2016). However, these systems do not provide computer-based as well as experiential knowledge-based support for CBL practice. In CBL practice, the clinical case is a key component, which provides a foundation to understand disease particulars and enables students to use their experiential knowledge to interpret them easily (Demircioğlu & Selçuk, 2016).

Regarding experiential knowledge-based support, we have already developed an *interactive case-based learning system* (iCBLS) (Ali, Han, Bilal et al., 2018) to utilize the strength of humans (experiential knowledge). The iCBLS lacked the support of machine-generated domain knowledge. Currently, much less attention is given to provide the support of domain knowledge while formulating the case. According to the study by Ali et al. (2018), "case formulation means identification of a medical chart's components (demographics, chief complaint, medical history, habits, family history, medicines, allergies, diagnosis, treatment, and recommendations) from a given clinical case and then writing personal observations for each component." To design an effective CBL approach for better clinical competency, the following major research question must be answered:

- How can the gaps between human-based learning and CBL be filled to innovate the CBL approach for better clinical proficiency? Humans and computers have strengths and weaknesses (Cummings, 2014; Halim, 2018). In the medical area, human (domain expert) judgement is considered more credible than a computer. However, a human cannot perform fast reasoning computations to work for extended periods and will get tired and feel bored. A computer has the advantage over a human of being able to perform fast reasoning computation without feeling bored. Currently, much less attention is given to fill the gaps between human-based learning and CBL. Therefore, designing and developing an interactive and effective CBL approach to utilize the strength of both humans (experiential knowledge) and computers (domain knowledge), and overcoming the limitations, is our main target.

Knowledge is the wisdom of information that plays an important role in decision-making (Ali, Lee, & Kang, 2016). There exists plenty of textual data in the medical domain that can be useful for medical education, especially for CBL purposes. This data is available in a variety of formats and with different semantics. This overwhelming data provides various opportunities to gain useful knowledge that reflects the depth of information that plays an important role in decision-making. Declarative knowledge (also called factual knowledge) is a type of knowledge expressed in the form of unstructured text, which can play an important role in health's education, decision support, and wellness applications after structured transformation (Ali et al., 2016). According to the Simply Philosophy study (Philosophy, 2018), "factual knowledge is a justified affirmation of something." It combines the concepts to make an affirmation of something. For example, "blood_disease" and "is a symptom" make an affirmation "blood_disease is a symptom." The produced affirmation is either true or false. However, in declarative knowledge, it is always true. Handling unstructured content is the foundation to construct the domain knowledge (structured declarative knowledge) required for interactive learning to prepare medical students for better clinical practice.

This declarative knowledge can play an important role in real-life applications for better analysis if the unprocessed text is transformed into structured content (i.e., explicit knowledge). A huge amount of valuable textual data is available on the web, which has led to a corresponding interest in technology for automatically extracting relative information from open data, to convert it into declarative knowledge, and to represent it in a way that is machine interpretable. One way to represent this knowledge is ontology, which represents a machine-readable reality using a restriction-free framework, where you can explicitly define, share, reuse, and or distribute information. Ontology has been considered as a common way to represent real-world declarative knowledge (Lee, Kao, Kuo, & Wang, 2007). The research community prefers to use natural language processing (NLP) techniques to construct machine-readable knowledge. In the literature, most systems/methodologies (Friedman, Shagina, Lussier, & Hripcsak, 2004; Leao, Revoredo, & Baiao, 2013; Rajni & Taneja, 2013) require high intervention of a knowledge engineer to translate unstructured text into a structured form and to resolve the construction of unambiguous machine-readable knowledge. For an automated CBL, a structured knowledge construction from textual data is a challenging task (Rusu et al., 2013). In the text mining domain, normally, text preprocessing, text transformation, feature selection, term extraction, relation extraction, and model construction tasks are involved.

Keeping in view these facts, we responded to these deficiencies by including a methodology called KCM-CD to construct machine-readable domain knowledge (i.e., structured declarative knowledge) from unstructured text to facilitate and provide machine-generated domain knowledge to medical students for solving real-world clinical cases during CBL practice. The KCM-CD methodology constructs an ontology from unstructured textual resources in a systematic and automatic way using artificial intelligence techniques with minimum intervention from a knowledge engineer. For effective transformation, controlled natural language is used, which constructs syntactically correct and unambiguous computer-processable texts. In addition, to select the important features for domain knowledge construction, the KCM-CD methodology applies our previously proposed *univariate ensemble-based feature selection* (uEFS) methodology (Ali, Ali, Kim et al., 2018), which is an efficient and comprehensive methodology to filter out irrelevant features from an input data set. Furthermore, the KCM-CD methodology covers all major phases of cross industry standard process for data mining (CRISP-DM) to explain the end-to-end knowledge engineering process (Ali, Ali, Khan et al., 2018). To realize the KCM-CD methodology, we enhanced our developed CBL system called iCBLS to utilize the strength of both human (experiential knowledge) and computer (domain knowledge). The iCBLS was designed based on current CBL practices in the *School of Medicine*, *University of Tasmania*, *Australia*. This study expands our previous work as described by Ali, Han, Bilal, et al. (2018), Ali et al. (2015, 2016) that lacked the support of machine-generated domain knowledge as well as features/concepts selection and had limited results.

The motivation behind this methodology is to construct domain knowledge from unstructured text, to facilitate and provide machine-generated domain knowledge to medical students for CBL rehearsal. To achieve this goal, this study was undertaken with the following objectives: (a) to fill gaps between human-based and computer-based learning to innovate the CBL approach for better clinical proficiency, (b) to construct an ontology from unstructured textual resources without involvement of a knowledge engineer, (c) to automate the ontology development process without requiring extensive training in knowledge engineering to reduce the human resource cost, and (d) to boost the development of machine-generated knowledge-based systems. Figure 1 shows the study overview and flow of the paper.

The key contribution of this paper is to introduce an automatic methodology for constructing domain knowledge with minimum intervention of a knowledge engineer to fill the gaps between human-based and computer-based learning for better clinical proficiency.

The study is organized as follows. Section 2 describes related works and Section 3 presents the architecture of the proposed knowledge construction methodology and functional mapping of the KCM-CD methodology to phases of CRISP-DM. Section 4 describes a case study to explain the process of the KCM-CD methodology and an overview of the interactive CBL System. Section 5 evaluates the developed model. Finally, Section 6 concludes the paper with a summary of research findings and future directions.

2 | RELATED WORKS

This section describes various existing studies related to each aspect of this research work. This research focuses on presenting a methodology for constructing a reliable domain knowledge to innovate the CBL approach. Therefore, this section presents an overview of the CBL approach and different methodological studies of domain knowledge construction. Various research directions related to (a) CBL methodologies and (b) technologies used for domain knowledge construction are discussed in each subsection.

2.1 | CBL literature

Case-based learning is an active learning approach that focuses on clinical, community, and scientific problems. In CBL, the facilitator has an active role and authentic cases for clinical practice are used (Thistlethwaite et al., 2012; Umbrin, 2014). The CBL approach is one of the successful approaches in student-based pedagogy and it is widely applied in medical education (Eseonu et al., 2013; Gade & Chari, 2013; Osinubi &

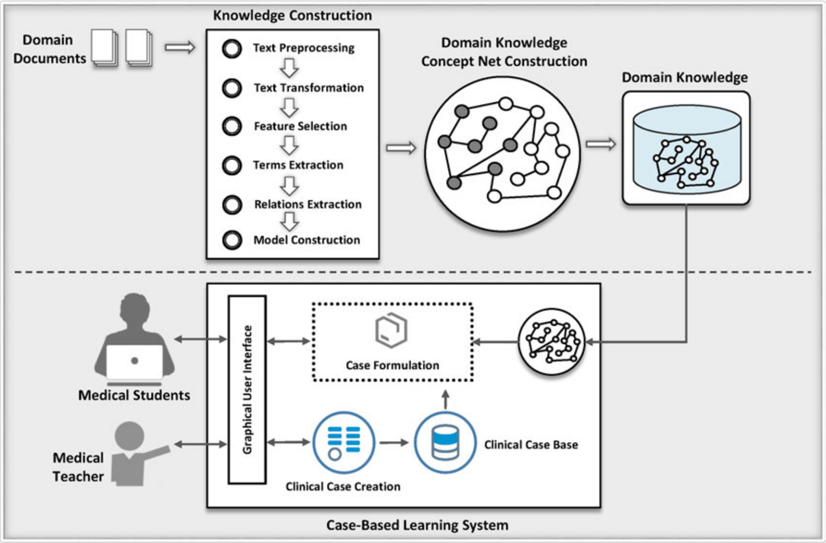


FIGURE 1 Idea diagram of the proposed research studies

Ailoje-Ibru, 2014; Patil & Karadesai, 2016; Sule, 2016). CBL has been used in clinical as well as non-clinical courses such as nursing courses, adult health, paediatric and obstetrical nursing courses, pathophysiology, statistics and research (Fish, 2005; Hoffman, Hosokawa, Blake Jr, Headrick, & Johnson, 2006).

In professional health care education, students tackle uncertain situations resulting from the accumulation of multiple problems (Baillergeau & Duyvendak, 2016). In such situations, students have their own judgement, opinion, and feedback and consider these integral and appropriate to the situation. Baillergeau and Duyvendak (2016) relate this situation with bricolage and investigated ways to correlate non-expert knowledge with other types of knowledge (expert knowledge). In such situations, experiential knowledge is considered a valuable resource (Baillergeau & Duyvendak, 2016; Popay & Williams, 1996) that can facilitate and provide lived knowledge to students to enhance an individual's participation and to empower users (Baillergeau & Duyvendak, 2016).

According to Willoughby and Philosophy (2018), ‘‘Experiential knowledge is a knowledge of particular things gained by perception and experience.’’ Similarly, Baillergeau and Duyvendak (2016) noted that ‘‘Experiential knowledge is a type of knowledge that has the potential to enhance the understanding of the nature, causes, and most effective responses to social problems.’’ Experiential knowledge either recalled from experiences, learned, or acquired (Storkerson, 2009) is mostly utilized for problem solving. Teachers, general practitioners, and social workers are the leading experts that provide experiential knowledge. These experts provide competent interventions utilizing their practical knowledge that is built up using experiential or lay knowledge. Experiential knowledge can be domain-specific as well as holistic and is mostly described in the form of statements (Storkerson, 2009). The idea of experiential expertise, introduced in early 1980s (Duyvendak, 1999). Willoughby and Philosophy (2018), observed that the brain has remarkable capacity for accumulating information and facts. She described that an older brain has accumulated and stored vastly more information than a younger brain. An older person has a well of information and experience to draw on. Therefore, age and experience are advantages in fields like coaching, journalism, law, and management. According to Storkerson (2009), ‘‘The term experience refers to the interactions that humans have with their environments.’’ Similarly, Baillergeau and Duyvendak (2016) stated that ‘‘practical knowledge is a key element in clinical knowledge and clinicians build this up through face-to-face observations, screening, and evaluation of persons.’’ Experiential knowing is an endless practice of perception and decision-making, which is an important aspect for analysing experiential knowledge (Storkerson, 2009).

In the medical area, human (domain expert) judgement is considered more credible than a computer. However, a human cannot perform fast reasoning computation to work for long periods and they experience fatigue and can feel bored (Rodriguez-Barbero & Lopez-Novoa, 2008). A computer has the advantage over a human in being able to perform fast reasoning computations, while not experiencing boredom.

2.2 | Domain knowledge construction literature

According to Abacha and Zweigenbaum (2011), ‘‘the medical knowledge is growing significantly every year. According to some studies, the volume of this knowledge doubles every 5 years, or even every 2 years.’’ Because most of the information available in digital format is unstructured (Feldman et al., 2002), the information extraction problem has attracted wide interest in several research communities (Doan, Ramakrishnan, & Vaithyanathan, 2006). Text mining is a multidisciplinary research area that derives high-quality information from textual data and includes information retrieval, NLP, data mining (DM), machine learning, and other techniques (Baitule & Chole, 2014). In the text mining domain, normally text preprocessing, text transformation, feature selection, term extraction, relation extraction, and model construction tasks are involved to construct domain knowledge from textual data. For reliable knowledge construction, keywords, as well as their relations are the key elements for knowledge representation, which are mostly extracted from given data using machine learning approaches and a thesaurus (Chen & Lin, 2010; Wenchao, Lianchen, & Ting, 2009). Loh, Wives, and de Oliveira (2000) noted that concept extraction is a low-cost process that helps to build a vocabulary for constructing/discovering domain knowledge. Haggag (2013) described that both qualitative and quantitative techniques can be used for keywords extraction task. Qualitative techniques are considered reliable, whereas quantitative techniques are preferable due to handling multiple text processing tasks. In the literature, various keyword extraction methodologies are used which are represented in Figure 2.

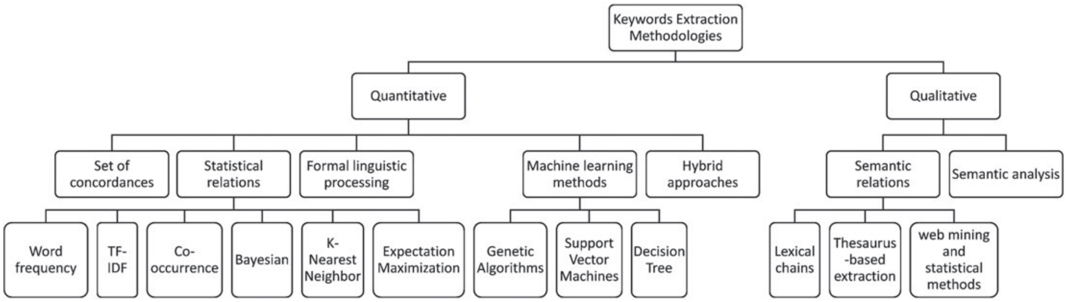


FIGURE 2 Keyword extraction methodologies (Abacha & Zweigenbaum, 2011; Azcarraga, Liu, & Setiono, 2012; Feng et al., 2011; Haggag, 2013)

TABLE 1 Advantages and disadvantages of technologies used for domain knowledge construction

Reference	Method/Technique/Tool	Advantages	Disadvantages
Liu and Wang (2007)	Corpus dependent approach for keyword extraction	– Provides better performance	– Requires documents and fixed keywords to develop a prediction model for single domain
Al-Khalifa and Davis (2006); Chen and Lin (2010); Beliga et al. (2014); Haggag (2013); Feng et al. (2011); Azcarraga et al. (2012)	Statistical approaches for keyword extraction	– Considered as simplest models – Used for Complex terms extraction	– Filter out important infrequent keyword, – Results have low precision, – Require hand-annotated data sets for learning, – Restricted by word-related features and become more complex by adding more features
Lott (2012); Robertson (2004); Wren et al. (2004); Al-Khalifa and Davis (2006)	Word Frequency Analysis -Term Frequency Inverse Domain Frequency (TF-IDF)	– Determines good candidate keywords, – Most commonly used due to having simplicity and effectiveness characteristics	– Does not always discover meaningful relationship between words, – Term frequency Ignores the contents' semantics
Feldman et al. (2002)	Co-occurrence strategy	– Constructs rules in fast manner, – Much simpler strategy to seek the relevant terms without syntactic or semantic consideration	– Relatively low precision
Abacha and Zweigenbaum (2011)	MetaMap for medical entity recognition	– Maps medical text to UMLS concepts for identifying the precise concepts	– Recognizes some common words as medical terms, – Considers multiple concepts and their semantic types for the same term, – Needs a disambiguation step to obtain required concept
Wenchao et al. (2009)	Naive Bayes technique	– Simple technique to produce good results	– Requires manually assigned keywords for model training
Abacha and Zweigenbaum (2011)	Domain-dependent relation extraction methods	– Extracts relations using meta-thesaurus and semantic network of UMLS	– Dependent on domain knowledge
Haggag (2013); Ercan and Cicekli (2007); Feng et al. (2011)	WordNet	– Enables to calculate the similarity between noun as well as verb pairs, – Provides semantic features within words	– Supports limited vocabulary and not covers all domains

Various technologies are used that help to construct domain knowledge from textual data. Each method, technique, or tool involved in the knowledge construction process has advantages and disadvantages, which are illustrated in Tables 1 and 2.

Rajni and Taneja (2013) proposed a framework called U-STRUCT that converts textual documents into an intermediate structured form. However, a knowledge engineer is required to convert that intermediate form into a fully structured form. Similarly, Friedman et al. (2004) developed an approach that maps textual data into UMLS codes for translating them into a structured form (XML format). However, their approach does not support lexical ambiguity and requires a knowledge engineer as well as domain knowledge for structured translation. Leao et al. (2013) proposed an ontology learning methodology using OntoUML. They converted unstructured text into structured form by utilizing full proposed and implemented a semi-automatic methodology to extract knowledge from unstructured as well as semi-structured data. The proposed methodology does not support lexical ambiguity.

Controlled natural languages are a subset of natural language which is easily understandable by humans (Kuhn, 2009). CNL is a restricted language that can be processed and interpreted by computers. This language preserves its essential properties while restricting its syntax, semantics, and lexicon (Kuhn, 2014). CNL was proposed to build knowledge bases (ontologies). Multiple CNLs have been developed to build semantic web ontologies such as *Attempto Controlled English* (ACE), *Sydney OWL Syntax* (SOS), *Controlled Language for Ontology Editing* (CLOnE), and *Rabbit*. In the literature, various categories of CNLs are used which are represented in Figure 3.

For computer processability, the CNL is written in a formal logic. The basic purpose of defining a CNL is to design computer-processable text for improving machine translation. Safwat and Davis (2014) noted that CNLs facilitate non-expert users to develop ontologies of varying sizes in an easy-to-use manner. Williams, Power, and Third (2014) described how CNLs are knowledge representation languages, which help non-expert users to translate their knowledge into a computer interpretable form without involvement of a knowledge engineer. In addition, Miyabe and Uozaki (2014) described various features of CNL, namely that they: (a) enhance readability, (b) improve terms disambiguity, (c) are easy to understand, (d) minimize the role of a knowledge engineer, (e) reduce the human translation cost, and (f) improve reusability of knowledge.

Kuhn (2009) designed a CNL, called Attempto Controlled English (ACE), which is considered one of the most mature CNLs. ACE was developed in early 1995 and has been under development for more than 20 years. This language is most widely used in the academic domain. Its vocabulary is not fixed and varies based on the problem domain. ACE also covers all four design principles, as compared with other CNLs that do not satisfy all principles. In addition, it is acknowledged to be an unambiguous language. Similarly, Denaux (2013) also described some features of the ACE language. He noted that ACE can be used for ontology construction without the knowledge of Web Ontology Language (OWL). It supports all kinds of ontology expressiveness. In addition, it is easy to use for all domain experts.

In the literature, most systems/methodologies (Friedman et al., 2004; Leao et al., 2013; Rajni & Taneja, 2013) require a knowledge engineer to translate unstructured text into fully structured form and most systems have been developed using NLP techniques and without the support of controlled natural language (Friedman et al., 2004; Houser, 2004; Jindal & Taneja, 2013). Regarding structured knowledge construction, some studies do not support lexical ambiguity (Rajni & Taneja, 2013; Reuss et al., 2015). We responded to these deficiencies by including a KCM-CD methodology to construct the domain knowledge (i.e., structured declarative knowledge) from unstructured text. For effective transformation, controlled natural language is used, which constructs syntactically correct and unambiguous computer-processable texts (Kuhn, 2009).

TABLE 2 Advantages and disadvantages of technologies used for domain knowledge construction (cont.)

Reference	Method/Technique/Tool	Advantages	Disadvantages
Loh et al. (2000)	Word-sense disambiguation	– Natural language processing techniques help to solve the ambiguity problems	– Requires complex algorithms and time to analyse the text, – Requires knowledge models and rules
Gazendam et al. (2010)	Restricted vocabulary or thesaurus	– Produce consistent results	– Construction and maintenance of thesaurus
Ercan and Cicekli (2007); Lott (2012)	Lexical chains	– Widely used in text summarization, – Locate terms and their sequence in quick and accurate manner	– Not fully explored in keyword extraction problems, – Is an exhaustive method

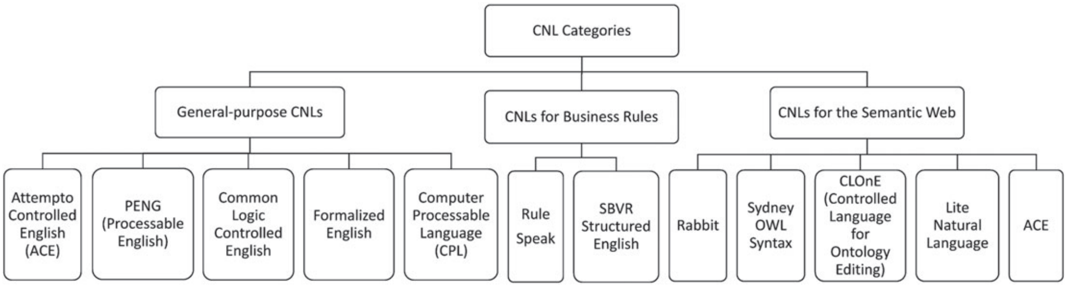

CNL Categories
General-purpose CNLs
CNLs for Business Rules
CNLs for the Semantic Web
Attempto Controlled English (ACE)
PENG (Processable English)
Common Logic Controlled English
Formalized English
Computer Processable Language (CPL)
Rule Speak
SBVR Structured English
Rabbit
Sydney OWL Syntax
CLOnE (Controlled Language for Ontology Editing)
Lite Natural Language
ACE

FIGURE 3 Controlled natural languages categories (Kuhn, 2009)

3 | MATERIALS AND METHODS

To construct a domain knowledge from textual data using text mining process, this section describes (a) the KCM-CD methodology and module details and (b) functional mapping of the KCM-CD methodology to the CRISP-DM phases.

3.1 | Proposed knowledge construction methodology

Text mining is the process of deriving high-quality information from an unstructured text. For constructing machine-readable domain knowledge from textual data, a workflow of the KCM-CD methodology is shown in Figure 4, which consists of six modules—*text preprocessing*, *text transformation*, *feature selection*, *terms extraction*, *relations extraction*, and *model construction*.

A brief description of each module is below.

3.1.1 | Text preprocessing

The text preprocessing module applies various basic preprocessing techniques to prepare textual data. This module consists of four components—*tokenization* for chopping the given text into pieces (tokens), *filtration* for removing the non-informative terms (such as the, in, a, an, with, etc.), *tagging* for assigning each token with a parts-of-speech tag such as noun, verb, etc., and *normalization* for identifying the root/stem of a word, i.e., the words "connected" and "connecting" are stemmed to "connect."

3.1.2 | Text transformation

This module computes the *term frequency—inverse document frequency* (TF-IDF) of the extracted tokens to generate feature vectors (tabular form) representing document instances.

3.1.3 | Feature selection

This module applies our previously proposed *univariate ensemble-based feature selection* (uEFS) methodology (Ali, Ali, Kim et al., 2018) to select the important features for domain knowledge construction. The uEFS methodology includes two innovative algorithms: (a) Unified features scoring algorithm to generate a final ranked list of features following a comprehensive evaluation of a feature set and (b) threshold value selection algorithm to define cutoff points for removing irrelevant features (see Ali, Ali, Kim, et al. (2018) for details).

3.1.4 | Terms extraction

A concept expresses more concrete and accurate meanings than keywords do. For identifying concept relationships and building domain ontology, there is need to extract concepts (i.e., named entities) from the given textual data. The terms extraction module configures an external thesaurus (i.e., Princeton's WordNet) to identify concepts by mapping all nouns of the processed textual data with the concepts defined in a thesaurus. This module is responsible for identifying relevant terms.

3.1.5 | Relations extraction

For generating a concepts hierarchy to build a domain ontology, identification of concept relationships is needed which can be achieved by using an external semantic lexicon. The relations extraction module extracts relations based on linguistic patterns using external semantic lexicons. This module performs the semantic analyses to define the meanings of words and their relations by mapping with domain-specific standard vocabularies. Finally, this module validates the concept relations from the domain expert before model construction.

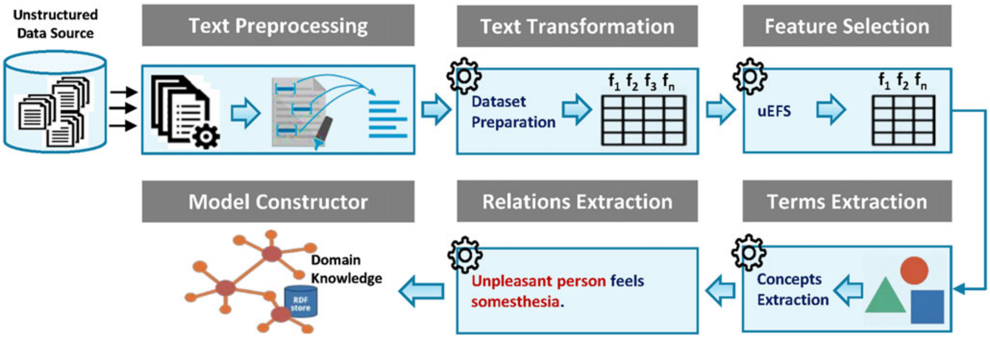


FIGURE 4 A workflow of the KCM-CD methodology

TABLE 3 Methods used for constructing domain knowledge

Process	Task	Method	Reason
Text preprocessing	Tokenization	English tokenizer	
	Filtration	Stopword removal	
	Normalization	Porters stemmer	
	Tagging	POS tagger	
Text transformation	Technique used	Term Frequency – Inverse Document Frequency (TF-IDF) Univariate Ensemble-based Features Selection (uEFS)	1. TF-IDF contributes a good heuristic for determining likely candidate keywords (Lott, 2012). 2. It is the most commonly used keyword extraction algorithm currently in use (Robertson, 2004) when a document corpus is available.
Feature selection	Technique used		uEFS is an efficient and comprehensive feature selection methodology to filter out the irrelevant features (Ali, Ali, Kim et al., 2018).
Terms extraction	Process	Nouns, Verbs, Adjectives, and Adverbs Identification	Penn Treebank (Taylor et al., 2003) provides distinct coding for all classes of words having distinct grammatical behaviour.
	Thesaurus used	Penn Treebank	
Relations extraction	Technique used	Lexical chaining and heuristics	Lexical chain is a well known technique for text connectivity (Ghosh, 2014) that locates terms and their sequence in an accurate manner (Lott, 2012).
	Thesaurus used	Princeton's WordNet	
	Process	Hypernyms identification	
	Keep original tokens	True	
	Multiple meanings per word policy	Take all meanings per token	
	Multiple synset words	Take only first synset word	
	Validation	Domain expert	
Model construction	Language used	Attempto Controlled English (ACE)	1. ACE (Kuhn, 2009) is a knowledge representation language, which utilizes the syntax of a subset of English. 2. It supports automatic and unambiguous translation of text into first-order logic.

TABLE 4 CRISP-DM phases and tasks performed in the KCM-CD methodology (Marbán, Mariscal, & Segovia, 2009)

Business understanding	Data understanding	Data preparation	Modelling	Evaluation	Deployment
Understand application domain	Search domain documents	Text tokenization	Select features	Evaluate the results of uEFS methodology	Plan deployment
Identify application goal	Collect initial documents	Remove stopwords	Extract terms	Evaluate the extracted terms	Monitor application impact
Identify application objectives	Analyse documents	Terms stemming	Extract relations	Evaluate the extracted relations	Maintain application
Analyse resource specification (software, hardware)	Remove irrelevant documents	POS tagging	Convert to ACE	Determine next steps	Prepare final report
Prepare application development plan	Store required documents	Text transformation	Construct model		Review application

TABLE 5 A partial view of feature vectors

Action	Agonist	.....	Blood	Bloodstream	BMI	.....	Label
0.0000	0.0044	.. …	0.0119	0.0000	0.0155	.. …	Diabetes
0.0020	0.0005	.. …	0.0510	0.0000	0.0079	.. …	Diabetes
0.0029	0.0204	.. …	0.0323	0.0025	0.0247	.. …	Diabetes
0.0009	0.0039	.. …	0.0306	0.0000	0.0000	.. …	Diabetes
0.0021	0.0008	.. …	0.0530	0.0000	0.0055	.. …	Diabetes
0.0025	0.0025	.. …	0.0816	0.0000	0.0066	.. …	Diabetes
0.0015	0.0042	.. …	0.0431	0.0000	0.0190	.. …	Diabetes
0.0016	0.0042	.. …	0.0437	0.0000	0.0192	.. …	Diabetes
0.0032	0.0023	.. …	0.0303	0.0000	0.0000	.. …	Diabetes
0.0013	0.0000	.. …	0.0000	0.0000	0.0000	.. …	Non-diabetes
0.0000	0.0000	.. …	0.0000	0.0000	0.0000	.. …	Non-diabetes
0.0007	0.0000	.. …	0.0013	0.0000	0.0000	.. …	Non-diabetes
0.0006	0.0000	.. …	0.0007	0.0000	0.0000	.. …	Non-diabetes
0.0010	0.0000	.. …	0.0000	0.0000	0.0000	.. …	Non-diabetes
0.0007	0.0000	.. …	0.0006	0.0000	0.0000	.. …	Non-diabetes
0.0017	0.0000	.. …	0.0021	0.0000	0.0000	.. …	Non-diabetes
0.0006	0.0000	.. …	0.0000	0.0000	0.0000	.. …	Non-diabetes
0.0022	0.0000	.. …	0.0019	0.0000	0.0000	.. …	Non-diabetes
0.0000	0.0000	.. …	0.0000	0.0000	0.0000	.. …	Non-diabetes

TABLE 6 Top diabetes domain words extracted from clinical documents

Diabetes Domain Words				
action	prevention	child	beverage	triglyceride
agonist	sick	cholesterol	BMI	unstable
antidiabetic	stage	dietary	mellitus	reduce
blood	type	eat	diagnose	condition
bodyweight	critical	education	diastolic	woman
chest	cycle	excretion	dietitian	adult
diabetes	drug	glucagon	episode	judgement
diabetic	energy	obese	fat	gestational
diet	external	overweight	foot	height
fatness	failure	plasma	glycemia	cough
glucose	food	pressure	haemoglobin	fatigue
glargine	goal	protection	hemoprotein	breakfast
hormone	healthy	urine	hospitalization	syndrome
insulin	level	complication	hypertension	vital
lifestyle	medication	exercise	injection	avoid
lower	substance	tired	intake	problem
monitor	yield	metformin	intensive	indicator
nutrition	activity	vision	habit	frequent
obesity	aged	hdl	goal	coma
visualize	influenza	hyperglycemia	disease	lispro
amount	adult	hypoglycemia	regular	hyper
walk	breathless	metabolic	pregnancy	thirst
drink	feet	protein	repeat	glimepiride
growth	person	weight	sugar	high
prevent	serum	training	systolic	loss

3.1.6 | Model construction

This module transforms the relations into an unambiguous and syntactically correct processable text using controlled natural language (CNL). The CNL helps to construct the structured ontological model called a domain model. As according to Denaux (2013) and Kuhn (2007), CNL can transform the textual data into machine interpretable knowledge and can consume less memory as well as computing power.

To construct domain knowledge, each above-mentioned module has performed some task(s) and used method(s) as illustrated in Table 3. For text preprocessing, text transformation, terms extraction, and relation extraction modules, the *RapidMiner Studio* was used (Mohammed, Mohammed, Fiaidhi, Fong, & Kim, 2014), whereas the *ACE View* was used for the model constructing module. The ACE View uses *Attempto Controlled English* to view and edit OWL ontology (Kaljurand, 2008).

3.2 | Functional mapping of the KCM-CD methodology with phases of the CRISP-DM

The CRISP-DM is a widely used systematic methodology for data science projects (Ali, Ali, Khan et al., 2018). CRISP-DM consists of six well-defined phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Shearer, 2000). The KCM-CD methodology constructs an ontology from unstructured textual resources in a systematic as well as an automatic way using artificial intelligence techniques and covers all major phases of CRISP-DM to explain the end-to-end knowledge engineering process (Ali, Ali, Khan et al., 2018). Table 4 illustrates functional mapping of the KCM-CD methodology to the phases of CRISP-DM and individual tasks performed by each phase.

4 | REALIZATION OF THE KCM-CD METHODOLOGY

This section describes a case study to explain the process of the knowledge construction methodology (KCM-CD) and an overview of the interactive CBL System.

TABLE 7 Selected words for domain model construction

Diabetes domain words along with their weights	
<weight name="blood" value="0.998"/>	<weight name="lispro" value="0.362"/>
<weight name="diabetes" value="0.998"/>	<weight name="hyper" value="0.362"/>
<weight name="diabetic" value="0.998"/>	<weight name="thirst" value="0.362"/>
<weight name="diet" value="0.998"/>	<weight name="glimepiride" value="0.362"/>
<weight name="glucose" value="0.998"/>	<weight name="high" value="0.305"/>
<weight name="glargine" value="0.998"/>	<weight name="loss" value="0.305"/>
<weight name="insulin" value="0.998"/>	<weight name="feeling" value="0.279"/>
<weight name="obesity" value="0.998"/>	<weight name="edema" value="0.273"/>
<weight name="level" value="0.751"/>	<weight name="tension" value="0.273"/>
<weight name="feet" value="0.743"/>	<weight name="unpleasant" value="0.273"/>
<weight name="person" value="0.743"/>	<weight name="negative" value="0.256"/>
<weight name="serum" value="0.743"/>	<weight name="symptom" value="0.231"/>
<weight name="pressure" value="0.743"/>	<weight name="negative_stimulus" value="0.194"/>
<weight name="metformin" value="0.743"/>	<weight name="blood_disease" value="0.165"/>
<weight name="vision" value="0.743"/>	<weight name="bloodpressure" value="0.123"/>
<weight name="hdl" value="0.743"/>	<weight name="somesthesia" value="0.123"/>
<weight name="hyperglycemia" value="0.743"/>	<weight name="blurry" value="0.123"/>
<weight name="weight" value="0.587"/>	<weight name="medicine" value="0.108"/>
<weight name="glycemia" value="0.587"/>	<weight name="feel" value="0.108"/>
<weight name="hypertension" value="0.587"/>	<weight name="swallow" value="0.105"/>
<weight name="disease" value="0.485"/>	<weight name="oat" value="0.060"/>
<weight name="regular" value="0.485"/>	<weight name="urination" value="0.059"/>
<weight name="fatigue" value="0.388"/>	<weight name="hurt" value="0.059"/>
<weight name="indicator" value="0.373"/>	<weight name="stimulus" value="0.059"/>
<weight name="frequent" value="0.373"/>	<weight name="salmon" value="0.050"/>
<weight name="coma" value="0.362"/>	<weight name="felt" value="0.050"/>

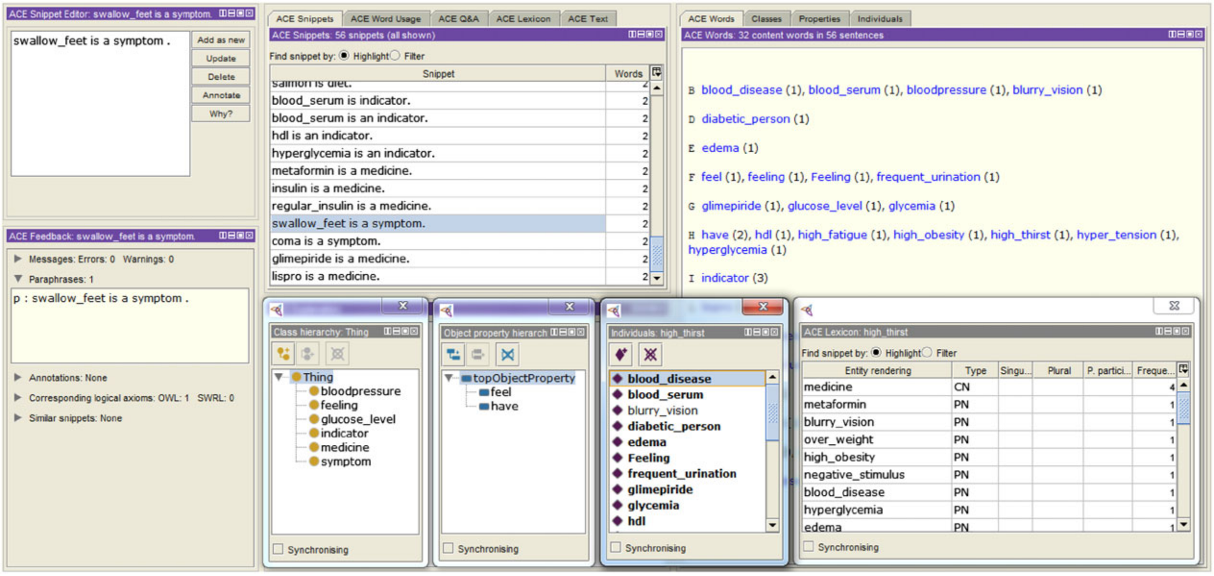
4.1 | Case study: diabetes

According to the Ali study (Ali, Han, Bilal et al., 2018), "case study helps to perform an in-depth study and analysis of a real-world or an imagined scenario." To explain the process of the KCM-CD methodology, a case study of the clinical documents of diabetes and non-diabetes domains is considered. The steps for realization of the KCM-CD methodology are:

1. Load the clinical documents of diabetes and non-diabetes domains.
2. Perform the text preprocessing task, including text tokenization, stopwords removal, tokens filtration, terms stemming, and POS tagging, on loaded documents.
3. Compute the TF-IDF of each term to generate the feature vectors for transforming the text into structured form as shown in Table 5.
4. Compute the ranks of each feature using proposed uEFS methodology and then select the important features (words) of diabetes domain only as shown in Table 6.
5. Extract terms (words) after identification of nouns, verbs, adjectives, and adverbs using Penn Treebank as shown in Table 7.
6. Extract and identify all entities relations using the lexical chain technique and a heuristic approach. For example, lexical chain extracts "*symptom/blood_disease*" and "*symptom/feeling/somesthesia/unpleasant_person/negative_stimulus/hurt* " (Ali et al., 2016) relations of "*symptom*" word.
7. Finally, for the model construction process, first transforms the identified relations into an unambiguous and syntactically correct processable text using controlled natural language (CNL) as shown in Table 8.
8. Write the transformed text into the ACE editor (see Figure 5) to construct the structured ontological model, also called the domain model, as shown in Figure 6. Once the ontological domain model is built, it can be accessed and used by medical students for better clinical competency (Baillergeau & Duyvendak, 2016).

TABLE 8 Identified relations of diabetes domain

Attempto Controlled English (ACE) text	
feeling is a symptom.	high_obesity is a symptom.
somesthesia is a feeling.	over_weight is a symptom.
unpleasant_person feels somesthesia.	edema is a symptom.
unpleasant_person has negative_stimulus.	blood_serum is an indicator.
negative_stimulus is a hurt.	hdl is an indicator.
blood_disease is a symptom.	hyperglycemia is an indicator.
glycemia is glucose_level.	metformin is a medicine.
hyper_tension is bloodpressure.	regular_insulin is a medicine.
weightlost is a symptom.	swallow_feet is a symptom.
frequent_urination is a symptom.	glimepiride is a medicine.
high_thirst is a symptom.	lispro is a medicine.
high_fatigue is a symptom.	glargine is a medicine.



swallow_feet is a symptom .

Find snippet by: ● Highlight ○ Filter

Snippet	Words
salmon is diet.	2
blood_serum is indicator.	2
blood_serum is an indicator.	2
hdl is an indicator.	2
hyperglycemia is an indicator.	2
metaformin is a medicine.	2
insulin is a medicine.	2
regular_insulin is a medicine.	2
swallow_feet is a symptom.	2
coma is a symptom.	2
glimepiride is a medicine.	2
lispro is a medicine.	2

B blood_disease (1), blood_serum (1), bloodpressure (1), blurry_vision (1)

D diabetic_person (1)

E edema (1)

F feel (1), feeling (1), Feeling (1), frequent_urination (1)

G glimepiride (1), glucose_level (1), glycemia (1)

H have (2), hdl (1), high_fatigue (1), high_obesity (1), high_thirst (1), hyper_tension (1), hyperglycemia (1)

I indicator (3)

▶ Messages: Errors: 0 Warnings: 0
▼ Paraphrases: 1
p : swallow_feet is a symptom .
▶ Annotations: None
▶ Corresponding logical axioms: OWL: 1 SWRL: 0
▶ Similar snippets: None

- Thing
  - bloodpressure
  - feeling
  - glucose_level
  - indicator
  - medicine
  - symptom

- topObjectProperty
  - feel
  - have

- blood_disease
- blood_serum
- blurry_vision
- diabetic_person
- edema
- Feeling
- frequent_urination
- glimepiride
- glycemia
- hdl

Find snippet by: ● Highlight ○ Filter

Entity rendering	Type	Singu...	Plural	P. partici...	Freque...
medicine	CN				4
metaformin	PN				1
blurry_vision	PN				1
over_weight	PN				1
high_obesity	PN				1
negative_stimulus	PN				1
blood_disease	PN				1
hyperglycemia	PN				1
edema	PN				1

FIGURE 5 Domain model generation through ACE-controlled natural language

4.2 | Interactive CBL system overview

We designed and developed the iCBLS, a system for medical education to practice real-world CBL cases (Ali, Han, Bilal et al., 2018). This system is a web-based application that enables medical educators to create real-world CBL cases for their students with the support of their experiential knowledge and computer-generated trends, review students' solutions, and give feedback and opinions to their students. It also facilitates medical students to do CBL rehearsal before attending an actual CBL class. The output of this system is the course's information, real-world cases, health records, formulated cases, and the teacher's feedback (Ali, Han, Bilal et al., 2018). The iCBLS was designed based on the current CBL practices in the *School of Medicine*, *University of Tasmania*, *Australia* (see (Ali, Han, Bilal et al., 2018) for details).

This study expands our previous work as mentioned in Ali, Han, Bilal, et al. (2018), Ali et al. (2015) that lacked the support of machine-generated domain knowledge. To realize the knowledge construction methodology, we enhanced our developed iCBLS to utilize the strength of both humans (experiential knowledge) and computers (domain knowledge) for better clinical competency. This research will allow medical students to do CBL rehearsal with machine-generated domain knowledge support before attending an actual CBL class.

The simulation of the iCBLS is illustrated in Figure 7, where the partial view of ontological model (domain model) is shown. Figure 7 depicts three sections of the interface. The first section provides the description of a real-world CBL case, whereas the second section allows medical students to add chart components and then loads the ontological model that enables medical students to view the domain knowledge to record their personal observations of each component during their CBL practice. Finally, the third section shows the list of students who already formulated that case. After formulating a CBL case, students submit their data to get feedback from their teachers.

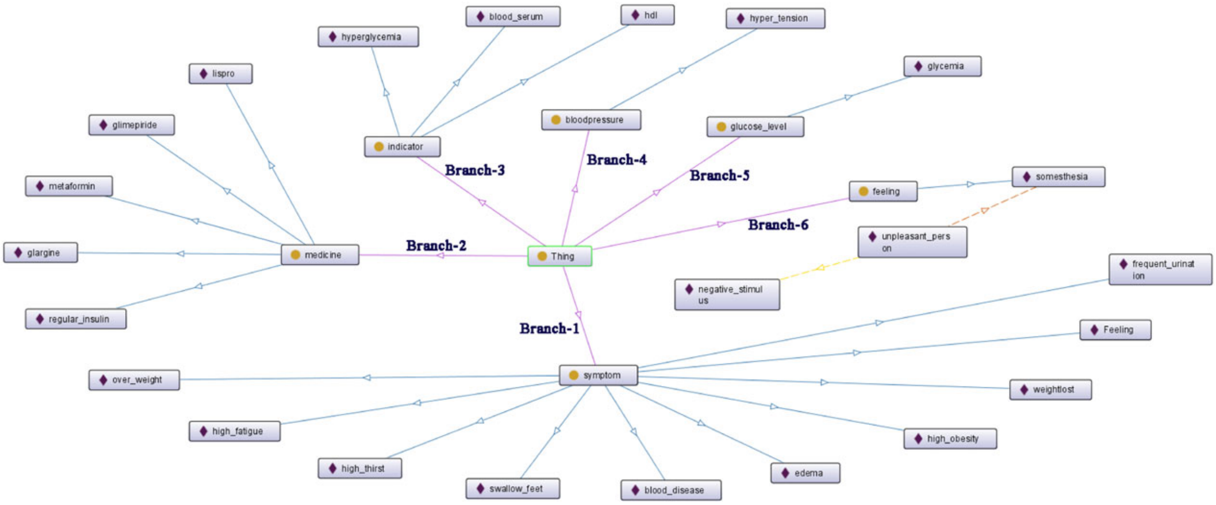
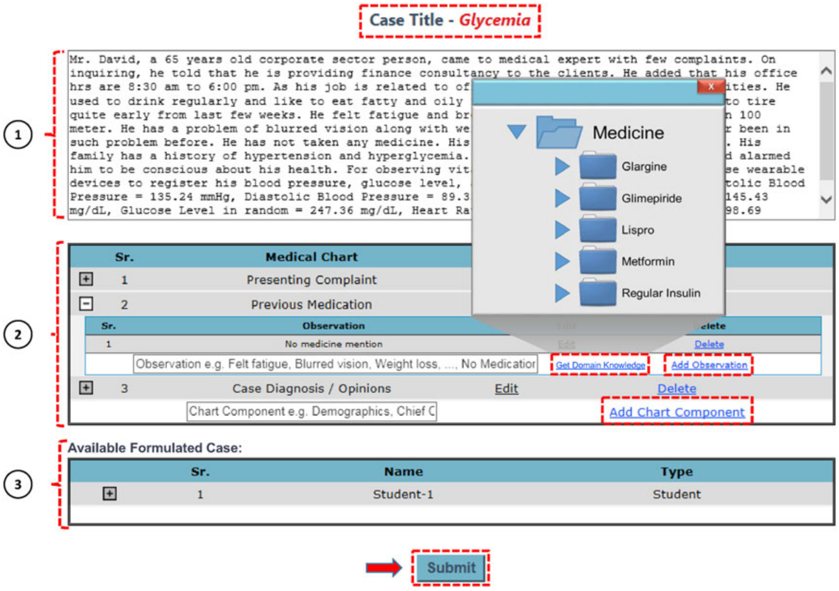


FIGURE 6 A partial view of the domain model



Case Title - *Glycemia*

```
Mr. David, a 65 years old corporate sector person, came to medical expert with few complaints. On
inquiring, he told that he is providing finance consultancy to the clients. He added that his office
hrs are 8:30 am to 6:00 pm. As his job is related to of                                  ities. He
used to drink regularly and like to eat fatty and oily                                   to tire
quite early from last few weeks. He felt fatigue and br                                  n 100
meter. He has a problem of blurred vision along with we                                  r been in
such problem before. He has not taken any medicine. His                                  . His
family has a history of hypertension and hyperglycemia.                                  d alarmed
him to be conscious about his health. For observing vit                                  se wearable
devices to register his blood pressure, glucose level,                                   tolic Blood
Pressure = 135.24 mmHg, Diastolic Blood Pressure = 89.3                                  145.43
mg/dL, Glucose Level in random = 247.36 mg/dL, Heart Ra                                  98.69
```

	Sr.	Medical Chart		
⊞	1	Presenting Complaint		
⊟	2	Previous Medication		
⊞	3	Case Diagnosis / Opinions	Edit	Delete

Sr.	Observation	Edit	Delete
1	No medicine mention	Edit	Delete

Observation e.g. Felt fatigue, Blurred vision, Weight loss, ..., No Medicatior | Get Domain Knowledge | Add Observation

Chart Component e.g. Demographics, Chief C | Add Chart Component

Available Formulated Case:

	Sr.	Name	Type
⊞	1	Student-1	Student

FIGURE 7 Simulation of the interactive case-based learning system

5 | MODEL EVALUATION

The evaluation phase of any methodology has a key role in investigating the value of the proposed method. This section evaluates the developed model. The purpose was to check the quality of the domain knowledge.

We considered C_v as the coherence measure for evaluating the quality of the model, which is considered as the best coherence measure for evaluation purposes (Röder, Both, & Hinneburg, 2015; Röder, 2017). C_v is based on a sliding window that uses normalized pointwise mutual information (NPMI) and the cosinus similarity (Röder, 2017) to retrieve co-occurrence counts for the given words. This measure has also been proven to correlate with human ratings.

Generally, the coherence measure C is the cross product of the four sets (Röder et al., 2015), which is defined as follows:

$$C = S * M * P * \Sigma \tag{1}$$

Where S, M, P, Σ represent the segmentation of word subsets, confirmation measure, word probabilities, and set of aggregation functions respectively.

C_v is a category of coherence measure, which is defined as follows:

$$C_v = S_{set}^{one} * \tilde{m}_{cos(nlr,1)} * P_{sw(110)} * \sigma_a \tag{2}$$

Where S_{set}^{one}, $\tilde{m}_{cos(nlr,1)}$, $P_{sw(110)}$, σ_a represent the segmentation method having N number of subset pairs, normalized log-ratio indirect confirmation measure (NPMI), word counts using a sliding window of size 110, and summarization method (arithmetic mean of all confirmations) respectively.

An equation for the S_{set}^{one} is given as follows:

$$S_{set}^{one} = \{(W', W^*) | W' = \{w_i\}; w_i \in W; W^* = W\} \tag{3}$$

Where W, W^*, $W^{'}$ represent the total word set, existence of the subset, and occurrence of the subset respectively (see details of C_v in Röder et al. (2015)).

TABLE 9 C_v coherence measure results of the domain model

Branch	Words	Coherence
Branch-1	diabetes, symptom, urination, obesity, edema, feet, weight, thirst, influenza, fatigue	0.600
Branch-2	diabetes, medicine, lispro, glimepiride, metaformin, glargine, insulin	0.531
Branch-3	diabetes, indicator, hyperglycemia, serum, hdl	0.640
Branch-4	diabetes, blood, pressure, hypertension	0.636
Branch-5	diabetes, glucose, level, glycemia	0.684
Branch-6	diabetes, feeling, somesthesia, unpleasant, person, negative, stimulus	0.441
Average Coherence		**0.589**

TABLE 10 Comparisons of state-of-the art coherence measures with the C_v coherence measure

Branch	C_P	C_{UCI}	C_{UMass}	C_{NPMI}	C_A	C_V
Branch-1	0.004	−2.989	−4.929	−0.061	0.139	**0.600**
Branch-2	−0.345	0.540	−6.551	0.066	0.264	**0.531**
Branch-3	−0.057	−4.026	−5.386	−0.127	0.133	**0.640**
Branch-4	0.861	3.717	−2.593	0.286	0.508	**0.636**
Branch-5	0.237	−0.175	−3.872	0.056	0.240	**0.684**
Branch-6	−0.075	−1.088	−2.902	−0.004	0.115	**0.441**
Average Coherence	0.104	−0.67	−4.372	0.036	0.233	**0.589**

C_P is computed using a sliding window (size = 70), a one-preceding segmentation of the top words, and the confirmation measure of Fitelson's coherence (Röder, Both, & Hinneburg, 2015). C_{UCI} is computed using a sliding window (size = 10) and the pointwise mutual information (PMI) of all word pairs of the given top words (Newman, Lau, Grieser, & Baldwin, 2010). C_{UMass} is computed using document co-occurrence counts, a one-preceding segmentation, and a logarithmic conditional probability as confirmation measure (Mimno, Wallach, Talley, Leenders, & McCallum, 2011). C_{NPMI} is computed using a sliding window (size = 10) and the normalized pointwise mutual information (NPMI) (Aletras & Stevenson, 2013). C_A is computed using a context window (size = 5), the NPMI, and the cosinus similarity (Aletras & Stevenson, 2013). C_V is computed using a sliding window (size = 110), the NPMI, and the cosinus similarity (Röder et al., 2015).

TABLE 11 Comparisons of state-of-the art knowledge construction methodologies with the KCM-CD methodology

Approach	Phases	Output type	NLP support	CNLP support	KE support	Knowledge execution
Domain modelling (Leao et al., 2013)	Text preprocessing, semantic annotation, supersenses extraction, semantic type mapping, mapping to OntoUML	Ontology	Yes	No	Yes	Full
Knowledge Extraction Workbench (Sauer & Roth-Berghofer, 2014)	Domain detection, web community selection, linked data repository, content mining, data retrieval, processing raw data, processed data, knowledge extraction, extracted knowledge, application in knowledge container, evaluation	Taxonomies	Yes	No	Yes	Full
Multi-agent System (Reuss et al., 2015)	Keyword extraction, synonyms and hypernyms, collocation extraction, vocabulary extension, similarity assessment, association identification, clustering and case impart, sensitivity analysis, consistency check, and feeedback	Taxonomies, rules	Yes	No	Yes	Full
U-STRUCT (Jindal & Taneja, 2013)	Text processing (text analysis phase), generalized intermediate form generation (text synthesis phase)	Concept-based form or relational tables (intermediate form of text)	Yes	No	Yes	Partial
KCM-CD	**Text processing, text transformation, feature selection, terms extraction, relations extraction, model construction**	**Ontology**	**Yes**	**Yes**	**No**	**Full**

NLP, natural language processing; CNLP, controlled natural language processing; KE, knowledge engineer.

For holistic understanding, each branch of the domain model (shown in Figure 6) was considered as one topic and all words in that branch were considered as input for measuring the coherence among these words. For calculating the coherence value of the word set, the Palmetto tool (Röder, 2017) was used. For computing the results of C_v, the window size 110 was used as this measure produced maximum coherence value at this window size (Röder et al., 2015). Table 9 illustrates the set of words in each branch of the model and their corresponding coherence value. The results show that all branches of the domain model have positive coherence values, which indicates that all words in each branch are correlated with each other. An average value was also computed, which is also a positive value, which shows that the quality of the developed model is acceptable.

In this study, we also considered other state-of-the art coherence measures, namely C_P, C_{UCI}, C_{UMass}, C_{NPMI}, and C_A (Aletras & Stevenson, 2013; Mimno, Wallach, Talley, Leenders, & McCallum, 2011; Newman, Lau, Grieser, & Baldwin, 2010; Röder et al., 2015) to evaluate the quality of the model. Table 10 illustrates the comparison of the C_v coherence measure of the developed model with state-of-the art other coherence measures. Other coherence measures are also explained along with the Table 10. The results show that the C_v coherence measure provides competitive results as compared with other state-of-the art coherence measures.

Finally, a comparison of the presented KCM-CD methodology with the state-of-the-art knowledge construction methodologies (Jindal & Taneja, 2013; Leao et al., 2013; Reuss et al., 2015; Sauer & Roth-Berghofer, 2014) was performed, which is illustrated in Table 11. All these methodologies convert unstructured text into a structured form to construct executable knowledge. All discussed methodologies in Table 11 do not have CNLP support whereas our proposed KCM-CD has this support. All other methodologies requires knowledge engineer support, but our case does not require it. This comparison shows that our proposed methodology outperforms all others in the domain of CBL, particularly in medical education. With the help of this model, we can automatically construct machine readable knowledge without the help of knowledge engineer, which helps medical students to learn the CBL case before the actual class.

6 | CONCLUSIONS AND FUTURE DIRECTIONS

In recent trends, more attention is given to e-learning environments for the clinical practice of medical students. To support learning outcomes, a large volume of web-based learning systems has been developed. However, most do not support computer-based interactive case formulation.

Medical literature contains a lot of useful knowledge in textual form, which can be beneficial for computer-based CBL practice. For an automated CBL, structured knowledge construction is a challenging task. Keeping in view these facts and to take care of the students' learning systems, this research investigated CBL and proposed a KCM-CD methodology to construct the ontological model from unstructured text. This will facilitate medical students to do CBL rehearsal with machine-generated domain knowledge support before attending an actual CBL class. This study expands our previous work (Ali, Han, Bilal et al., 2018; Ali et al., 2015) that lacked the support of machine-generated domain knowledge. We enhanced our developed iCBLS to utilize the strength of both humans and computers.

Domain knowledge can serve a broad range of applications such as decision support systems as well as education, health, and wellness applications. With the evolution of domain knowledge stored in a database, the developed CBL system can hold better clinical competence and provide intensive learning in the future. Currently, the proposed CBL approach does not support an interactive question–answering technique. In the future, we will extend the current CBL approach towards a QA-based (question–answer) learning environment.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support programme (IITP-2017-0-01629) supervised by the Institute for Information & Communications Technology Promotion (IITP). This work was supported by an IITP grant funded by the Korean government (MSIT) (IITP-2017-0-00655). This work was supported by the Industrial Strategic Technology Development Program, 10052955, Experiential Knowledge Platform Development Research for the Acquisition, and Utilization of Field Expert Knowledge, funded by the Ministry of Trade, Industry & Energy (MI, Korea).

The authors would like to thank Mrs. Kerrin McKeown for reviewing and editing the English in the manuscript. They would also like to express their sincere thanks to Hafiz Syed Muhammad Bilal and Muhammad Asif Razzaq for their valuable input to this manuscript. Lastly, they would like to express their deep gratitude to anonymous reviewers for their insightful comments to improve the quality of the paper.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Maqbool Ali was the principal researcher of this study. He devised and carried out the idea, designed and performed the experiments, and wrote the paper. *Jamil Hussain* contributed to formal analysis and model evaluation. *Sungyoung Lee* and *Byeong Ho Kang* provided advisory comments, remarks, and financial aid for the paper. *Kashif Sattar* revised and improved the quality of paper.

ORCID

Maqbool Ali https://orcid.org/0000-0002-4107-7122

REFERENCES

Abacha, A. B., & Zweigenbaum, P. (2011). Automatic extraction of semantic relations between medical entities: A rule based approach. *Journal of biomedical semantics*, *2*, S4.

Al-Khalifa, H. S., & Davis, H. C. (2006). Folksonomies versus automatic keyword extraction: An empirical study. *IADIS International Journal On Computer Science And Information Systems (IJCSIS)*, *1*, 132–143.

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, Potsdam, Germany, pp. 13–22.

Ali, M., Ali, R., Khan, W. A., Han, S. C., Bang, J., Hur, T., Kim, D., Lee, S., & Kang, B. H. (2018). A data-driven knowledge acquisition system: An end-to-end knowledge engineering process for generating production rules. *IEEE Access*, *6*, 15587–15607.

Ali, M., Ali, S. I., Kim, D., Hur, T., Bang, J., Lee, S., Kang, B. H., & Hussain, M. (2018). uEFS: An efficient and comprehensive ensemble-based feature selection methodology to select informative features. *PloS one*, *13*, e0202705.

Ali, M., Bilal, H. S. M., Hussain, J., Lee, S., & Kang, B. H. (2015). An interactive case-based flip learning tool for medical education. In *Inclusive Smart Cities and e-Health*, Springer, Geneva, Switzerland, pp. 355–360.

Ali, M., Han, S. C., Bilal, H. S. M., Lee, S., Kang, M. J. Y., Kang, B. H., Razzaq, M. A., & Amin, M. B. (2018). iCBLS: An interactive case-based learning system for medical education. *International Journal of Medical Informatics*, *109*, 55–69.

Ali, M., Lee, S., & Kang, B. H. (2016). uDeKAM: A methodology for acquiring declarative structured knowledge from unstructured knowledge resources. In *In Machine Learning and Cybernetics (ICMLC), 2016 International Conference on*, *1*, IEEE, Jeju, South Korea, pp. 177–182.

Azcarraga, A., Liu, M. D., & Setiono, R. (2012). Keyword extraction using backpropagation neural networks and rule extraction. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, IEEE, Brisbane, QLD, Australia, pp. 1–7.

Baillergeau, E., & Duyvendak, J. W. (2016). Experiential knowledge as a resource for coping with uncertainty: Evidence and examples from the Netherlands. *Health, Risk & Society*, *18*, 407–426.

Baitule, P., & Chole, V. (2014). A review on improved text mining approach for conversion of unstructured to structured text. *International Journal of Computer Science and Mobile Computing*, *3*, 156–159.

Beliga, S., Meštrović, A., & Martinčcić-Ipšić, S. (2014). Toward selectivity based keyword extraction for croatian news. *arXiv preprint arXiv*, *1407*, 4723.

Boubouka, M. (2013). A web-based case-based learning environment–Use in the Didactics of Informatics. (Ph.D. thesis), National and Kapodistrian University of Athens.

Chen, L.-S., Cheng, Y.-M., Sheng-Feng, W., Yong-Guo, C., & Lin, C.-H. (2009). Applications of a time sequence mechanism in the simulation cases of a web-based medical problem-based learning system. *Journal of Educational Technology & Society*, *12*, 149–161.

Chen, P.-I., & Lin, S.-J. (2010). Automatic keyword prediction using google similarity distance. *Expert Systems with Applications*, *37*, 1928–1938.

Cheng, Y.-M., Sheng-Huang, K., Shi-Jer, L., & Ru-Chu, S. (2012). The effect of applying online PBL case system to multiple disciplines of medical education. *TOJET: The Turkish Online Journal of Educational Technology*, *11*, 283–294.

Cummings, M. M. (2014). Man versus machine or man + machine? *IEEE Intelligent Systems*, *29*, 62–69.

Demircioğlu, S., & Selçuk, G. S. (2016). The effect of the case-based learning method on high school physics students' conceptual understanding of the unit on energy. In *Asia-Pacific Forum on Science Learning and Teaching, vol. 17, 1–25. The Education University of Hong Kong, Department of Science and Environmental Studies*, Hong Kong.

Denaux, R. (2013). Intuitive ontology authoring using controlled natural language. University of Leeds.

Doan, A., Ramakrishnan, R., & Vaithyanathan, S. (2006). Managing information extraction: State of the art and research directions. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, ACM, Chicago, IL, USA, pp. 799–800.

Duyvendak, J. W. (1999). De planning van ontplooiing: Wetenschap, politiek en de maakbare samenleving. Sdu.

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing & Management*, *43*, 1705–1714.

Eseonu, O., Carachi, R., & Brindley, N. (2013). Case-based anatomy teaching: A viable alternative? *The Clinical Teacher*, *10*, 236–241.

Feldman, R., Aumann, Y., Finkelstein-Landau, M., Hurvitz, E., Regev, Y., & Yaroshevich, A. (2002). A comparative study of information extraction strategies. *Computational Linguistics and Intelligent Text Processing*, LNCS 2276. Springer, Berlin, Heidelberg, pp. 349–359.

Feng, J., Xie, F., Hu, X., Li, P., Cao, J., & Wu, X. (2011). Keyword extraction based on sequential pattern mining. In *Proceedings of the third international conference on internet multimedia computing and service*, ACM, Chengdu, China, pp. 34–38.

Fish, T. T. T. (2005). If we teach them to fish: Solving real nursing problems through problem-based learning. *Annual Review of Nursing Education*, *3*, 109. Strategies for Teaching, Assessment, and Program Planning.

Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, *11*, 392–402.

Gade, S., & Chari, S. (2013). Case-based learning in endocrine physiology: An approach toward self-directed learning and the development of soft skills in medical students. *Advances in physiology education*, *37*, 356–360.

Gazendam, L., Wartena, C., & Brussee, R. (2010). Thesaurus based term ranking for keyword extraction. In *Database and Expert Systems Applications (DEXA), 2010 Workshop on*, IEEE, Bilbao, Spain, pp. 49–53.

Ghosh, A. (2014). Bengali Text Summarization using Singular Value Decomposition. (Ph.D. thesis), Jadavpur University.

Gopalan, C. (2016). The impact of rapid change in educational technology on teaching in higher education. *HAPS Educator*, *20*, 85–90.

Haggag, M. H. (2013). Keyword extraction using semantic analysis. *International Journal of Computer Applications*, *61*, 1–6.

Halim, S. (2018). Human and computer. Available online: https://www.comp.nus.edu.sg/~stevenha/viz/appendixC_hci.pdf Accessed: 2018-02-03.

Hoffman, K., Hosokawa, M., Blake Jr, R., Headrick, L., & Johnson, G. (2006). Problem-based learning outcomes: Ten years of experience at the university of missouri? Columbia school of medicine. *Academic Medicine*, *81*, 617–625.

Houser, A. (2004). Framemaker: Structured or unstructured?http://doi.org/www.writersua.com/articles/frame/

Jindal, R., & Taneja, S. (2013). U-struct: A framework for conversion of unstructured text documents into structured form. In *In Advances in Computing, Communication, and Control*, Springer, Berlin, pp. 59–69.

Kaljurand, K. (2008). ACE view—An ontology and rule editor based on attempto controlled english. In OWLED.

Kuhn, T. (2007). Authoring tools for ACE. http://doi.org/attempto.ifi.uzh.ch/site/docs/authoring_tools.html

Kuhn, T. (2009). Controlled English for knowledge representation. (Ph.D. thesis), Citeseer.

Kuhn, T. (2014). A survey and classification of controlled natural languages. *Computational Linguistics*, *40*, 121–170.

Leao, F., Revoredo, K., & Baiao, F. (2013). Learning well-founded ontologies through word sense disambiguation. In *2013 Brazilian Conference on Intelligent Systems*, Fortaleza, Brazil, pp. 195–200.

Lee, C.-S., Kao, Y.-F., Kuo, Y.-H., & Wang, M.-H. (2007). Automated ontology construction for unstructured text documents. *Data & Knowledge Engineering*, *60*, 547–566.

Liu, J., & Wang, J. (2007). Keyword extraction using language network. In *Natural language processing and knowledge engineering, 2007. NLP-KE 2007. International Conference on*, IEEE, China, pp. 129–134.

Loh, S., Wives, L. K., & de Oliveira, J. P. M. (2000). Concept-based knowledge discovery in texts extracted from the web. *ACM SIGKDD Explorations Newsletter*, *2*, 29–39.

Lott, B. (2012). Survey of keyword extraction techniques. *UNM Education*, *50*, 1–11.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. *Data Mining and Knowledge Discovery in Real Life Applications*, *2009*, 8.

McLean, S. F. (2016). Case-based learning and its application in medical and health-care fields: A review of worldwide literature. *Journal of Medical Education and Curricular Development*, *3*, JMECD–S20377.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, Honolulu, Hawaii, pp. 262–272. Association for Computational Linguistics.

Miyabe, M., & Uozaki, H. (2014). Controlled natural language simplifying language use. In *LREC2014Workshop-CNL Proceedings*, pp. 1–2.

Mohammed, O., Mohammed, S., Fiaidhi, J., Fong, S., & Kim, T. (2014). Clinical narratives context categorization: The clinician approach using rapidminer. *International Journal of Bio-Science and Bio-Technology*, *6*, 45–56.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 Annual Conference of the North American chapter of the association for computational linguistics*, Los Angeles, California, pp. 100–108. Association for Computational Linguistics.

Osinubi, A. A., & Ailoje-Ibru, K. O. (2014). A paradigm shift in medical, dental, nursing, physiotherapy and pharmacy education: From traditional method of teaching to case-based method of learning-a review. *Annual Research and Review in Biology*, *4*, 2053–2072.

Patil, M. J., & Karadesai, S. (2016). To determine the effectiveness of case based tutorials as compared to traditional tutorials in microbiology. *National Journal of Integrated Research in Medicine*, *7*, 5–8.

Philosophy, S. (2018). Factual knowledge. Available online: https://doi.org/simplyphilosophy.org/study/factual-knowledge/. Accessed: 2018-02-03.

Popay, J., & Williams, G. (1996). Public health research and lay knowledge. *Social science & medicine*, *42*, 759–768.

Rajni, J., & Taneja, S. (2013). U-STRUCT: A framework for conversion of unstructured text documents into structured form. *Advances in Computing, Communication, and Control*, 59–69.

Reuss, P., Althoff, K.-D., Henkel, W., Pfeiffer, M., Hankel, O., & Pick, R. (2015). Semi-automatic knowledge extraction from semi-structured and unstructured data within the omaha project. In *International Conference on Case-Based Reasoning*, Springer, Cham, pp. 336–350.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for idf. *Journal of documentation*, *60*, 503–520.

Röder, M. (2017). Palmetto online demo. Available online: http://doi.org/palmetto.aksw.org/palmetto-webapp/. Accessed: 2018-07-02.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, ACM, Shanghai, China, pp. 399–408.

Rodriguez-Barbero, A., & Lopez-Novoa, J. (2008). Teaching integrative physiology using the quantitative circulatory physiology model and case discussion method: Evaluation of the learning experience. *Advances in Physiology Education*, *32*, 304–311.

Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M., & Marinescu, V. (2013). Converting unstructured and semi-structured data into knowledge. In *Roedunet International Conference (RoEduNet), 2013 11th*, IEEE, Sinaia, Romania, pp. 1–4.

Safwat, H., & Davis, B. (2014). A brief state of the art of cnls for ontology authoring. In *International Workshop on Controlled Natural Language*, Springer, Galway, Ireland, pp. 190–200.

Sauer, C. S., & Roth-Berghofer, T. (2014). Extracting knowledge from web communities and linked data for case-based reasoning systems. *Expert Systems*, *31*, 448–456.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of data warehousing*, *5*, 13–22.

Shyu, F.-M., Liang, Y.-F., Hsu, W., Luh, J.-J., & Chen, H.-S. (2004). A problem-based e-learning prototype system for clinical medical education. *Medinfo*, *11*, 983–987.

Storkerson, P. (2009). Experiential knowledge, knowing and thinking. In *Experiential Knowledge Method & Methodology: International Conference*, pp. 1–18.

Suebnukarn, S., & Haddawy, P. (2007). Comet: A collaborative tutoring system for medical problem-based learning. *Intelligent Systems IEEE*, *22*, 70–77.

Sule, R. (2016). Medical students and zaculty perceptions towards a case-based learning Intervention at an Indian Medical College, McMaster University, Ph.D. thesis.

Taylor, A., Marcus, M., & Santorini, B. (2003). The penn treebank: An overview, Treebanks, pp. 5–22.

Thistlethwaite, J. E., Davies, D., Ekeocha, S., Kidd, J. M., MacDougall, C., Matthews, P., Purkis, J., & Clay, D. (2012). The effectiveness of case-based learning in health professional education. A beme systematic review: Beme guide no. 23. *Medical Teacher*, *34*, e421–e444.

UNM (2016). Extension for community healthcare outcomes - echo, the University of New Mexico. http://doi.org/echo.unm.edu/. Accessed: 2016-12-04.

UTMB (2013). Design a case, University of Texas Medical Branch - UTMB. http://doi.org/www.designacase.org/default.aspx. Accessed: 2016-12-04.

Umbrin, I. (2014). Difference between problem based learning PBL and case-based learning CBL, 2014. http://doi.org/www.slideshare.net/izzaumbrin/difference-between-problem-based-learning-pbl-and-case-based-learning-cbl. Accessed: 2017-01-21.

Wenchao, M., Lianchen, L., & Ting, D. (2009). A modified approach to keyword extraction based on word-similarity. In *Intelligent computing and intelligent systems, 2009. ICIS 2009. IEEE international conference on*, *3*, IEEE, Shanghai, China, pp. 388–392.

Williams, S., Power, R., & Third, A. (2014). How easy is it to learn a controlled natural language for building a knowledge base? In *International Workshop on Controlled Natural Language*, Springer, pp. 20–32.

Willoughby, D., & Philosophy, S. (2018). Experiential knowledge. Available online: https://doi.org/simplyphilosophy.org/study/experiential-knowledge/. Accessed: 2018-02-03.

Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., & Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, *20*, 389–398.

AUTHOR BIOGRAPHIES

Maqbool Ali received the MSc degree (Hons) in computer science from the University of Agriculture at Faisalabad in 2001 and the MS degree (Hons) in IT from the National University of Sciences and Technology, Islamabad, Pakistan, in 2013. Currently, he has completed the conjoint PhD degree with the Department of Computer Science and Engineering, Kyung Hee University, South Korea, and the School of Engineering and ICT, University of Tasmania, Australia. He has over 10 years of experience in the IT industry. He has contributed to over 30 publications in various reputed journals/conferences. His research interests include knowledge engineering, machine learning, natural language processing, and case-based learning.

Jamil Hussain received the MS degree from the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Pakistan, in 2014. He is currently pursuing the PhD degree with the Ubiquitous Computing Laboratory, Department of Computer Engineering, Kyung Hee University, South Korea. He has a professional experience of over 7 years in software industry working on user experience design and development on various projects. His research interest includes user experience design, artificial intelligence, and information extraction from textual data.

Sungyoung Lee received the BS degree from Korea University, Seoul, South Korea, and the MS and PhD degrees in computer science from Illinois Institute of Technology, Chicago, IL, USA, in 1987 and 1991, respectively. He was an Assistant Professor with the Department of Computer Science, Governors State University, University Park, IL, USA, from 1992 to 1993. He has been a Professor with the Department of Computer Engineering, Kyung Hee University, South Korea, since 1993, where he has been the Director of the Neo Medical ubiquitous-Life Care Information Technology Research Center since 2006. He is currently the Founding Director of the Ubiquitous Computing Laboratory. His current research interests include ubiquitous computing and applications, wireless ad hoc and sensor networks, context-aware middleware, sensor operating systems, real-time systems and embedded systems, and activity and emotion recognition. He is a member of ACM.

Byeong Ho Kang received the PhD degree from the University of New South Wales, Sydney, in 1996. He was a visiting researcher with the Advanced Research Laboratory, Hitachi, Japan. In addition, he has played a role in the foundation of several spinoff companies. He is currently a professor at the School of Engineering and ICT, University of Tasmania, Australia. He leads the Smart Services and Systems Research Group of post-doctoral scientists, which has carried out fundamental and applied research in expert systems, web services, SNS analysis, and smart industry areas. He has been involved in the development of several commercial and Internet-based applications, including AI products, expert system development tools, intelligent help desk systems, web-based information monitoring and classification systems, and so on. His research interests include basic knowledge acquisition methods and applied research in Internet systems and medical expert systems. He has served as the chair and a steering committee member in many international organizations and conferences.

Kashif Sattar received his MS degree in Information Technology in 2011 from National University of Sciences and Technology (NUST), Islamabad, Pakistan. He completed his PhD in Information Technology from NUST in 2017. Currently, he is working as a lecturer and project coordinator at Arid University Rawalpindi, Pakistan. He worked on PingER project in collaboration with SLAC, Stanford University. Also he got Junior Associate award from International Centre for Theoretical Physics (ICTP), Italy. His research interests include wireless mesh networks, optimization modelling, PCB routing, network performance analysis, algorithm analysis, and network security.

How to cite this article: Ali M, Hussain J, Lee S, Kang BH, Sattar K. A knowledge construction methodology to automate case-based learning using clinical documents. *Expert Systems*. 2019;e12401. https://doi.org/10.1111/exsy.12401