Research on Integrating Explicit/Implicit Semantic Representation and Multimodal Knowledge Graph for Traditional Chinese Medicine Digital Therapy

Longqing Zhang, Lei Yang, Xinwei Zhang*, Yungui Chen, Yongjian Huang and Jiawei Zhan
Guangdong University of Science and Technology, DongGuan, China

Keywords: Integrating Explicit, Implicit Semantic, Chinese Medicine Digital Therapy.

Abstract: The application of Artificial Intelligence (AI) technology is well-suited for Traditional Chinese Medicine (TCM) due to its reliance on observation through "looking, smelling, questioning, and cutting", as well as empirical diagnosis utilizing images, sounds, pulse sensing data, and other factors. This makes TCM an important area for breakthroughs in AI technology. The primary goal of this project is to extract a large quantity of TCM diagnostic knowledge that can be read by computers, train the TCM knowledge map model to become a discriminative model, and allow the model to differentiate between pairs of entities with different relationships or identify meaningful pairs of entities selected from randomly sampled negative entities. Constructing the TCM knowledge graph involves three main modules: TCM knowledge extraction, TCM knowledge fusion, and TCM knowledge computation. TCM knowledge extraction involves identifying the constituent elements of the knowledge graph, such as entities, relationships, and attributes, from vast amounts of semi-structured, structured, or unstructured pharmaceutical data, and determining the most effective method for depositing these elements into the knowledge base. TCM Knowledge Fusion integrates, disambiguates, and processes the contents of the TCM knowledge base, enhancing the logic and expressiveness within the knowledge base, and updating outdated knowledge or supplementing new knowledge for the TCM knowledge graph.

1 INTRODUCTION

Among the many application industries of AI technology, Chinese medicine is an important breakthrough direction in the application of AI technology because of its empirical diagnosis through images, sounds, and pulse sensing data in the way of "looking, smelling, questioning, and cutting", which is naturally compatible with the application characteristics of AI technology. On the basis of data-based diagnostic technology of Chinese medicine characteristics, widely incorporating modern medical micro-indicators, using big data and artificial intelligence methods, exploring new methods of diagnosis and classification of diseases, can better establish intelligent diagnosis decision support system with Chinese characteristics (Sun Z, 2018).Therefore, there is an urgent need to establish a real-world clinical research paradigm in Chinese medicine, regardless of the identification and treatment, or treatment effects, complex paradigm, using Chinese medicine clinical-based big data for Chinese medicine research, clinical research integration, building Chinese medicine structured electronic medical records, improving Chinese medicine clinical information collection system, building Chinese medicine literature and clinical database, developing Chinese medicine big data and artificial intelligence technology applications, so as to promote the great development of TCM.

This project proposes the research topic of "Research on TCM digital therapy integrating explicit/implicit semantic representation and multimodal knowledge mapping", which is to research and develop an intelligent diagnosis system for TCM, covering the data of TCM diagnosis, medicines, and cases by focusing on the big data of TCM diagnosis and treatment in the field of Artificial Intelligence + TCM, AI technology, and Knowledge Mapping technology, and build a platform for integrating science, industry, and education into one.

Zhang, L., Yang, L., Zhang, X., Chen, Y., Huang, Y. and Zhan, J.
Research on Integrating Explicit/Implicit Semantic Representation and Multimodal Knowledge Graph for Traditional Chinese Medicine Digital Therapy.
DOI: 10.5220/0012277900003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 214-219
ISBN: 978-989-758-677-4

2 RELATED WORK

Currently, the field of medicine stands as one of the extensively employed vertical domains for knowledge graph application. It is also a prominent research area in the realm of artificial intelligence, both domestically and internationally. The utilization of knowledge graphs in intelligent medical sectors, such as intelligent triage(McInerney J, 2018), disease risk assessment, intelligent assisted diagnosis and treatment, medical quality control, and medical knowledge, holds promising prospects for development. In China, numerous research teams have been actively utilizing artificial intelligence and knowledge graphs for training and exploration in this field.

Mi et al (Lu Y, 2018) utilized polynomial logistic regression (Logistic Regression), Random Forest (Random Forest), Support Vector Machine (Support Vector Machine), k-Nearest Neighbor (k-Nearest Neighbor), Decision Tree (Decision Tree), Decision Tree), Artificial Neural Network (ANN) and other machine learning algorithms to build prediction models for commonly used prescriptions and evaluated them. The possibility of prescription prediction and the amount of data required for robust prediction are elucidated. It is a comprehensive baseline model exploration of prescription recommendation applications.

Audema et al (Xuan P, 2019) utilized NLP (Natural Language Processing) and information mining techniques to make a remarkable contribution to the emergence of the first edition of our medical atlas, which was constructed to cover diseases, drugs, and diagnostic and therapeutic techniques, including more than 1 million instances of medical conceptual relationships.

Alshahrani (Perozzi B, 2014) and others used a meta-path-like randomized wandering strategy and performed the construction of input features and performed the recommendation task. Some researchers used heterogeneous network embedding representations for relationship prediction studies.

Collobert et al (Xiao D, 2023) used CNN (Convolutional Neural Network) model for named entity recognition; Chiu &Osama et al. combined CNN model with CRF model and designed residual expansion convolutional neural network RDCNN-CRF (Reduced Deep Convolutional Neural Network).

Overall, the current key tasks for knowledge graphs in healthcare are mainly focused on healthcare knowledge extraction and knowledge fusion. As for TCM healthcare, there are mostly domestic studies and fewer foreign ones, while knowledge graphs fusing explicit/implicit semantic representations and multimodality for TCM digital therapies are even more rarely seen.

3 SYSTEM DESIGN

This project introduces a "Study of Chinese Medicine Digital Therapy Incorporating Explicit/Implicit Semantic Representation and Multimodal Knowledge Mapping." Its system architecture, as depicted in Figure 1, revolves around leveraging big data related to Chinese medicine diagnosis and treatment within the realm of AI combined with traditional Chinese medicine. AI technology, and knowledge mapping technology, and researches and develops an intelligent diagnosis system for Chinese medicine, which covers the data of Chinese medicine diagnosis and treatment, medicines, and cases, and establishes a platform for the integration of science, industry, and education.

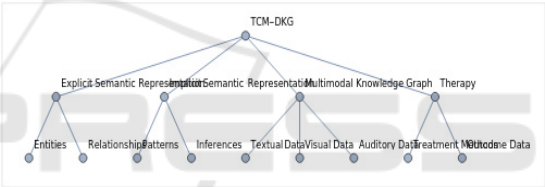

Figure 1: System architecture.

3.1 Knowledge Ontology Construction

The data schema, i.e., the ontology, is first defined in the construction of the knowledge graph, which is generally compiled manually by domain experts. Definition starts from the topmost concepts and then is gradually refined to form a well-structured hierarchy of categorized concepts. Under the guidance of the ontology, factual knowledge mining is performed on existing data sources to form a domain knowledge graph through entity discovery, relationship extraction, knowledge fusion and quality control.

The knowledge graph is populated with information derived from various sources, including structured, semi-structured, and unstructured data. To analyze and utilize this data, knowledge extraction techniques are applied to extract structured information that can be comprehended and processed by computers from the different structures and types of data.Knowledge acquisition is to extract knowledge from data of different sources and structures(Mikolov, 2013), to form

structured knowledge and deposit it into the knowledge graph. Currently, knowledge acquisition is mainly carried out for text data, and the extraction problems that need to be solved include: entity extraction, relationship extraction, attribute extraction and event extraction.

3.2 Domain-Specific Ontology Construction

Chinese medicine is a complex and huge system with thousands of types of entities, attributes, and relationships, and it is obvious that to build a complete knowledge rest system, it is far from enough to rely only on the power of expert manpower. For this reason, the automatic discovery capability of ontology needs to be vigorously studied. In the iterative process, the project uses the existing ontology as a guide, and applies weakly-supervised and unsupervised learning, such as remote supervision and clustering, to explore the general generalization and classification laws between factual knowledge (entities, and their attributes, and relationships) and conceptual knowledge (concepts, and their attributes, and relationships), so as to discover new ontologies, and concepts.

3.3 Evaluation and Naming of Basic Ontopsychological Concepts

The formation of basic mental concepts is influenced by a number of factors, the most important of which are the types and quantitative constraints on conceptual connotations. By connotation, we mean the attributes of the concept and their values. Connotation constraints, on the other hand, refer to the constraints on the range of values of attributes, which have the properties of commonness, ease of understanding, and so on. Connotation constraints and their evaluation laws can be learned from the mapping mechanism of existing ontological concepts and facts. The maximum entropy regression formula for concept evaluation can be expressed as:

$$e(c) = \frac{\prod_{p \in c} f(p)}{Z} \quad (1)$$

C is the target concept to be evaluated, which consists of multiple feature cluster constraints p with "or" relationships. Each feature cluster constraint p consists of multiple sub-feature constraints with "with" relationships. The sub-feature constraints are binary (attribute, attribute value range). If an attribute is constrained to take only one value, then the attribute value range is that value. If the metric perspective of this attribute is important, but the attribute value is not important (i.e., when the attribute needs to be considered qualitatively in the formation of a concept, but a specific measure is not needed, the range of values is noted as NULL). f(p) takes the value 1 only if all sub-feature constraints in p are satisfied, otherwise it is 0. Alternatively, p can be an overall measure of the feature constraints, e.g., the number of constraints, the ease of comprehension due to the structure, etc. Z is a normalization factor in order to get the evaluation value in the interval (0,1), which may not be computed in the selection of the best concept (Qu, 2023).

4 REMOTE SUPERVISED AUTOMATIC LABELING ALGORITHM

The lower layer of the model is common across all datasets, while the upper layer (specifically, CRF) produces outputs that are specific to each dataset. The character-level layer receives sentences from the dataset as input and captures contextual information at the character level using a BiLSTM, which produces representation vectors for the characters. These character-level vectors are then combined with word-level vectors and passed through a word-level BiLSTM. This generates a contextual representation that encompasses both word-level and character-level information. This shared representation is trained using our multi-task objective function. Finally, the CRF component of the model produces annotations for the input utterances based on the dataset it belongs to. We train separate multi-task learning models for each dataset.

4.1 Shared Layer

The input data of our dataset is represented as s={w1, w2, ⋯, wn}, where wi represents the ith word. To obtain word embeddings, we utilize a word-level embedding layer that takes the input sentence s and produces embeddings X = {x1, x2, ⋯, xn}. For character-level embeddings, we introduce a space character on both sides of each word to indicate the character input as c={c0,_, c1,0, ⋯, c1,_, c2,0, ⋯, cn, }, where ci,j denotes the jth character of the word wi in ci, and _ represents the

space character immediately following wi. Then, we map the individual characters in the sentence to character embeddings, denoted as C={c0,_, c1,0, ⋯, c1,_, c2,0, ⋯, cn}.

The character-level BiLSTM receives C as input and generates alternative representations for each word by concatenating the hidden vectors in the character space that follow the words in both forward and backward directions. It's important to note that initial word embeddings are obtained from a preprocessed word embedding lookup table based on a large corpus, while character embeddings are randomly initialized. During model training, both types of embeddings are fine-tuned.

The word-level Bi\LSTM takes the concatenated vectors of word embeddings and character-generated word vectors as input, generating final word representation vectors that effectively capture both word-level and character-level features. This framework allows the model to learn patterns based on characters and handle out-of-vocabulary (OOV) words, which are words not present in the word embedding lookup table, while still making full use of word embeddings.

4.2 Dataset-Specific CRF Layers

The chain-structured CRF is an effective framework for constructing probabilistic models of sequence labels that take into account the dependencies between sequence labels. Therefore, we built a dataset-specific CRF layer for sequence tag prediction (e.g., Bio NER and POS tagging in our experiments). We chose the IOBES tagging scheme for BioNER. The final word representation vector output from the shared layer is fed into the CRF component to generate sentence annotations y = {y1, y2, ⋯, yn}.

4.3 Small Sample Learning for Knowledge Graphs

Current approaches to knowledge graph complementation mainly map entities and relationships to a low-dimensional vector space, but utilize only the ternary structure <s,r,o> data in the knowledge graph, ignoring the text, pictures and numerical information that exist in large quantities in the knowledge base. This project proposes to embed the knowledge graph complementation model based on multimodal Linked Data. Embedding ternary as well as multimodal data together into the vector space not only makes link prediction more accurate, but also generates multimodal data with missing entities in the knowledge base to realize the knowledge graph complementation. Among them, the vector embedding of multimodal data is represented as follows:

a. Structured data: for entities of the knowledge base mapping, their unique heat codes are passed through a dense layer to get their embeddings;

b. Text: for those very short texts, such as names and titles, the characters are encoded using bi-directional GRUs; for those relatively long texts, the final encoding is obtained by CNN convolution and pooling over word vectors.

c. Images: using the VGG network pre-trained in the corpus, the embedding of the images is obtained

d. Numerical information: fully connected network, i.e., through a mapping that obtains numerical embeddings

The objective function for training is denoted as:

$$\sum_{(s,r,o)} \sum_{o} t_o^{s,r} \log\left(p_o^{s,r}\right) + \left(1 - t_o^{s,r}\right) \log\left(1 - p_o^{s,r}\right) \quad (2)$$

If the ternary <s, r, o> exists in the knowledge graph, the $t_o^{s,r}$ value is 1, otherwise it is 0. $p_o^{s,r}$ is the probability that this triad holds as predicted by the <s, r, o> model, which has a value between 0 and 1.

To tackle the aforementioned issues within the meta-learning framework, we propose a meta-learning algorithm that addresses target preference and under-emergence problems. Firstly, we incorporate an interactive attention extraction module as an additional component to enhance feature extraction. This module improves the distinguishability of feature vectors, mitigates the model's bias towards specific targets, and enhances its ability to generalize to novel tasks. Secondly, we employ graph neural networks to fully leverage the relationships among samples, constructing graph structures, and performing image classification at the node level. This approach significantly enhances the accuracy of classification by better capturing the inherent connections within the data.

5 QUESTION AND ANSWER TEXT MATCHING TECHNIQUES FUSING EXPLICIT AND IMPLICIT SEMANTIC REPRESENTATION

The implicit chapter relationship analysis task is actually a classification task, so the evaluation metrics are also commonly used for classification

tasks. The commonly used evaluation metrics are accuracy and decay value.

Accuracy is defined as:

$$Acc = \frac{|x : x \in X \wedge \overline{y} = y|}{|X|} \tag{3}$$

Where X is the total test data, |X | is the test data size, $\overline{y}$ is the true category of the corresponding data and y is the predicted category. Accuracy is the proportion of all correct predictions to the total data. The calculation of the decay value depends on the precision rate and recall rate. Precision rate is defined as:

$$P = \frac{|x : x \in X \wedge \overline{y} = y = label|}{|x : x \in X \wedge \overline{y} = label|} \tag{4}$$

Where x is the current category of interest. Accuracy is how much of all the data predicted for that category is correctly predicted. Words in natural language can be viewed as discrete symbols that cannot be given directly as input to a neural network, so a representation needs to be found to convert them into numerical inputs that the neural network can accept. In addition to vocabulary, there are many features that are also discrete symbols and also require a representation to be used as input. One traditional approach is to use solo heat vectors. This is done by first collecting all the required words to obtain a W. The unique heat vector of a word is then represented as an N-dimensional vector: viyone-hot[0,,,,,1,,,,0]T The ⁱth digit in this vector is 1, and all other digits are 0. The dimension of the unique heat vector is the same as the size of the word list, and thus for very large word lists, the dimensionality of the unique heat vector is extremely high, making it difficult to practically employ. This makes it difficult to adopt in practice, and also the fact that the unique heat vector uses different dimensions to represent different words makes the vector almost completely incapable of reflecting the semantic information of the words, and the fact that the representation of any two words is completely orthogonal makes it difficult to reflect the semantic information in this representation.

The primary procedure of our multilevel semantic fusion model proceeds as follows: Initially, we employ the SDT-CNN model proposed by us to acquire the representation of the implicit factual affective sentence SI found in the document D. Subsequently, within the factual implicit sentiment sentence SI, we consider the subject nouns as the objects of the sentiment target. Then, we adopt the FREERL model to extract the corresponding attributes associated with each object. The average of all object and attribute word embeddings is then employed as the representation of the sentiment target. In addition, all explicit sentiment sentences Sj ∈ SE within document D are treated as the contextual semantic background. We utilize a rule-based approach to categorize the sentiment polarity of these sentences. We learn the representation of each sentence Sj using the proposed SDT-CNN model, while the CNN-based model is used to learn the contextual semantic context representation. Finally, we combine the learned multilevel feature representations to form a comprehensive feature representation for classification.

6 SUMMARY

This topic is oriented to the knowledge extraction and fusion of Chinese medicine digital medical data, in the knowledge extraction stage, for unstructured data, based on its own irregular structural attributes can be obtained using deep learning techniques to obtain the relationship between the entities; for the structured information with certain rules or semi-structured form of the data, to take the crawler and parser to the entity relationship between the extraction and reconstruction. Knowledge fusion of the knowledge obtained from the two data sources mainly accomplishes the task of entity alignment, and then the TCM digital medical knowledge is stored to form a knowledge map, which provides data support for the downstream data application system.

ACKNOWLEDGMENTS

This research was financially supported by Special Projects in Key Areas for General Universities in Guangdong ProvinceNO.2021 ZDZX1077,in part of Natural Science Foundation of Guangdong Province of China with the Grant No.2020A1515010784,also supported by Guangdong Institute of Science and Technology Quality Project Editor GKZLGC2022255, 2022 Guangdong Institute of Science and Technology Innovation and Improvement School Project No. GKY-2022CQTD-2,2022 Guangdong Province Ordinary Colleges and Universities Young Innovative Talents Category Project, No. 2022KQNCX115,Innovation and Improvement School Project from Guangdong University of Science and Technology NO. GKY-2019CQYJ-3 College Students Innovation Training

Program held by Guangdong University of Science and Technology NO.1711034, 1711080, and NO.1711088.

REFERENCES

Sun Z,Yang J,Zhang J.Recurrent knowledge graph embedding for effective recommendation[C]. *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018: 297-305. https://dl.acm.org/doi/10.1145/3240323.3240361

McInerney J,Lacker B,Hansen S.Explore,exploit,and explain:personalizing explainable rec-ommendations with bandits[C]. *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018:31-39. https://dl.acm.org/doi/10.1145/3240323.3240354

Lu Y,Dong R,Smyth B.Why I like it:multi-task learning for recommendation and explanation[C]. *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018: 4-12. https://dl.acm.org/doi/10.1145/3240323.3240365

Xuan P,Cao Y,Zhang T.Dual Convolutional Neural Networks With Attention MechanismsBased Method for Predicting Disease-Related lncRNA Genes[J]. *Frontiers In Genetics*, 2019, 10(416): 1-1. https://doi.org/10.3389/fgene.2019.00416

Perozzi B,Al-Rfou R,Skiena S S.DeepWalk:Online Learning of Social Representations[C]. *In KDD'14:Proceedings of the 20th ACM SIGKDD International Conference on Knowledgediscovery and Data Mining*, 2014: 701-710. https://doi.org/10.48550/arXiv.1403.6652

Xiao D, Wenjun X U, Jiayi L I U, et al. Manufacturing capability service recommendation based on knowledge representation learning for industrial cloud robotics[J]. *Computer Integrated Manufacturing System*, 2023, 29(3): 719. http://www.cims journal.cn/EN/10.13196/j.cims.2023.03.003

Mikolov T,Chen K,Dean J.Distributed Representations of Words and Phrases and their Com-positionality[C]. *In Proceedings of the 26th International Conference on Neural Information Processing Systems*. New York, NY, USA, 2013: 3111-3119. https://doi.org/10.48550/arXiv.1310.4546

Bhat A D, Acharya H R, HR S.A Novel Solution to the Curse of Dimensionality in Using KNNsfor Image Classification[C]. *In 2019 2nd International Conference on Intelligent Autonomous Systems(ICoIAS)*, 2019: 32-36. https://www.doi.org/10.1109/icoias.2019.00012

Qu Y, Ma L, Ye W, et al. Towards Privacy-Aware and Trustworthy Data Sharing Using Blockchain for Edge Intelligence[J]. *Big Data Mining and Analytics*, 2023, 6(4): 443-464. https://www.doi.org/10.26599/bdma.2023.9020012