Knowledge Extraction From National Standards for Natural Resources: A Method for Multi-Domain Texts

Taiyu Ban, University of Science and Technology of China, China*
Xiangyu Wang, College of Data Science, University of Science and Technology of China, China
Xin Wang, Department of Computer Science and Technology, University of Science and Technology of China, China
Jiarun Zhu, Department of Computer Science and Technology, University of Science and Technology of China, China
Lvzhou Chen, College of Data Science, University of Science and Technology of China, China
Yizhan Fan, College of Data Science, University of Science and Technology of China, China

https://orcid.org/0000-0002-6509-6977

ABSTRACT

National standards for natural resources (NSNR) plays an important role in promoting efficient use of China’s natural resources, which sets standards for many domains such as marine and land resources. Its revision is difficult since standards in different domains may overlap or conflict. To facilitate the revision of NSNR, this paper extracts structural knowledge from the NSNR files to assist its revision. NSNR files are in multi-domain texts, where the traditional knowledge extraction methods could fall short in recalling multi-domain entities. To address this issue, this paper proposes a knowledge extraction method for multi-domain texts, including sub-domain relation discovery (SRD) and domain semantic features fusion (DSFF) module. SRD splits NSNR into sub-domains to facilitate the relation discovery. DSFF integrates relation features in the conditional random field (CRF) model to improve the capability of multi-domain entity recognition. Experimental results demonstrate that the proposed method could effectively extract structural knowledge from NSNR.

KEYWORDS

Entity Recognition, Hierarchical Clustering, Knowledge Extraction, Relation Discovery, Word Embedding

1. INTRODUCTION

National Standards for Natural Resources (NSNR) (Huang, C. et al., 2018; Wang, S. et al., 2018) plays an important role in the utilization of China’s natural resources such as marine, land and urban resources. It contains the standards of many kinds of important industries¹ and guides the future of
national development and natural resource utilization. NSNR sets standards for multiple domains (Huang, C. et al., 2018), and the standards of each domain are organized from the perspective of different disciplines², as shown in Figure 1. This characteristic could lead to duplicated and conflicting standards and makes the revision of NSNR difficult. Therefore, the knowledge extraction of NSNR (Wu, X et al., 2015) has significance for the analysis and revision of NSNR³.

For domain knowledge extraction (Chen, Y., 2018), entity recognition (Li, B., 2019) and relationship extraction (Wu, W. et al., 2019) are essential steps. The traditional methods of domain entity recognition and relationship extraction (Wang, B. et al., 2019) mainly include rule-based domain expert system (Jun-Ke, Z. et al., 2019) and machine learning (Chen, H. et al., 2008, 2010a) based methods (Chen, Y. et al., 2020; Liu, Z., &Chen, H., 2017). These methods mostly target certain domains with single industry or subject, instead of multiple domains. For the expert system, it has higher accuracy, but it is greatly time-consuming and requires high labor cost (Eftimov, T. et al. 2017). Moreover, the rule designing process is complicated, where various detailed features of domain entities and relationships should be considered requiring lots of manual labor. In addition, this kind of method has poor portability, an expert system is only applicable to one domain. When the domain changes, the expert system needs to be redesigned. For machine learning-based methods or some mixed methods (Jiang, Y., 2019; Savova, G. K. et al. 2010), they are more automatic (Qi, Y. et al., 2020) than the expert system. But these methods usually rely heavily on tagged corpus or domain relational database (Wang, J. et al. 2019). For NSNR, there are little tagged corpus. Therefore, it is difficult to use methods that demand lots of labeled domain corpus.

In this context, knowledge extraction from NSNR has the following challenges and features: (1) NSNR is a cross-industry domain (Liu, J. et al., 2019). It includes the standards of various industries, which contain multi-domain entities and relations (Mei, W., 2019; Yang, X. et al. 2019), leading to difficulties of domain entity and relation extraction. (2) For the standards of an industry (Jun-Lin, W., 2018; Zhang, Y., 2020), they are interdisciplinary which means that these standards are described in the perspective of various disciplines, as shown in Figure 1. So, the subjects among the same industry standard files are usually very different. (3) There lacks tagged corpus by experts for NSNR. (4) The entity relationships in NSNR tend to express recommendations or constraints in semantics, such as “should” and “must”. In response to the above features and challenges, designing a multi-dimension featured domain knowledge extraction method (Huang, X., 2017) which less depends on tagged corpus is of great significance for knowledge extraction from NSNR.

Traditional methods have limitations for the knowledge extraction from NSNR due to the pre-mentioned features. Due to the cross-industry characteristics of NSNR, the domain ontology designing becomes difficult. Moreover, the knowledge extraction methods that depend on the entity database of one certain domain also falls short due to the lack of tagged corpus. Besides, a sub-domain could include multiple industries, which makes it not feasible to split sub-domains based on the theme of industry. To solve these problems, this paper proposes a knowledge extraction method for multi-domain texts, which includes sub-domain relation discovery (SRD) module for relation extraction and domain semantic feature fusion (DSFF) module for entity recognition. For SRD, an indicator SDs_Db is designed to correctly split sub-domains from NSNR. A feedback mechanism integrates the indicator to iteratively discover the optimal splitting of the sub-domains. For DSFF, a conditional random field (CRF) based model integrates the domain relation features and semantic structures to improve the recall of multi-domain entities. It could recognize more entities through the representative relations (due to the feature (4) of NSNR) and the unique semantic patterns that are learned from a small set of tagged data.

The rest of this paper is organized as follows. Section II introduces previous works for domain knowledge extraction. Section III presents the overall framework of the proposed method. Section IV introduces the Sub-Domain Relation Discovery Model. Section V introduces the entity recognition model with domain semantic feature fusion. Section VI shows the experimental results and analyses. Section VII summarizes the contributions of this paper. Section VIII concludes this paper.
2. BACKGROUND

The focus and themes of domain knowledge extraction vary from domain to domain (Kwan, M. & Cheung, P., 2012; Chua et al., 2012). In recent years, many domain knowledge extraction methods (Liu et al., 2021; Chen et al., 2021; Netti et al., 2021) have been proposed in various industries, most of them focus on specific domains but less on multi-dimensional ones. Zhao et al. (2019) use the rule description method to recognize medical entities in unstructured medical data. They designed the patterns of medical entities and extracted the labels of entities through semantic analysis, which improved the accuracy of entity recognition in the medical domain. However, the method in this article focuses on the rule design of the lexical level of medical terms, which requires a great demand for experts’ domain analysis and not portable for other domains. Lin (2019) proposes a dependency tree-CRF based model to automatically extract patent information terms. And an improved k-means algorithm (Yao, Y., & Chen, H., 2018) is proposed to perform clustering analysis (Soares, R. G. F., Chen, H., & Yao, X, 2017) on term tags to obtain the category characteristics of entities. The method tends to guarantee the retrieval accuracy of the patent filed. It realizes the automation of feature mining through label analysis of domain entities, but the model lacks the filtering of prediction. The effect is strongly dependent on the recognition ability of the CRF model. Zhao et al. (2020) design a domain entity feature learning method based on CBOW and CEW. It takes full account of the multi-word composition of professional terms of medical entities. Through entity feature fusion RNN model, they improve the recognition accuracy of medical entities. Qi et al. (2020) fuses the military domain ontology features into word vectors through the analysis and design of domain ontology, enriching the domain information of the vector through the method of expert design. And they use BiLSTM-CRF model to train the domain feature-fused word vector, improving the accuracy of entity recognition in the military domain. Chen et al. (2020) utilizes a noise reduction learning method of the bidding domain entity recognition based on negative example data learning. They use the remote supervision
method of automatically labeling data in the field of bidding and combined the features of the field of bidding, realizing the entity relationship extraction in the field of bidding.

These methods mainly target domains which have single subjects. But they are not effective in multi-dimension featured domains, which consist of various industries or subjects. For knowledge extraction from multi-domain texts, Huang (2017) proposes a multi-dimensional semantic knowledge fusion framework for government website information analysis. It focuses on the domain ontology construction in multi-dimension featured data set. The framework uses clustering methods to study the relationships among various domains and realizes the ontology construction of the “Haze Prevention and Cure” domain. However, it is too complex for the lightweight multi-domain knowledge extraction and requires lots of labor on domain ontology designing (Kwahk et al., 2007). The single-subject domain targeted knowledge extraction methods are not adapted with multi-domain texts, which focus on the entities and relationships in only one subject. And existing methods for multi-domain texts mostly have a high cost on domain ontology designing.

3. OVERALL FRAMEWORK

This section briefly introduces the main components of the proposed method, i.e., modules of sub-domain relation discovery (SRD) and domain semantic features fusion (DSFF), as shown in Figure 2.

SRD consists of three parts which includes feedback based sub-domain clustering, sub-domain relationship clustering and sub-domain relation extraction, which are briefly introduced as follows:

1) Sub-domain clustering is designed to cluster the theme of multi-domain documents, which split the documents into sub-domains based on the similarity distances among documents. However, the number of total sub-domain clusters is unknown yet. In response, a feedback mechanism is proposed, which adjusts the critical distance of hierarchical clustering (Liu, X. et al., 2017) according to the effect of sub-domain relationship clustering. Moreover, an indicator $SDs_Dbw$ is designed to measure the effect of sub-domain relationship clustering. On this basis, a feedback mechanism adjusts the sub-domain clustering iteratively to produce the optimal sub-domain clustering (detailed in Section III).

2) Sub-domain relationship clustering performs the relation discovery in sub-domains. Dependency parsing (Jian, Z. et al. 2020) is executed on the sub-domain’s texts to obtain the dependency trees of texts. And word embedding (Nikfarjam et al., 2015; Segura-Bedmar et al. 2015; Xu, J. et al., 2016) is performed on the Chinese word segmentation (Feng, L. et al., 2020) of the dependency trees. Certain components of the dependency tree are extracted as the relation words. Subsequently, the relation representations are clustered and filtered to obtain the sub-domain’s relation trigger words.

3) Relation trigger word matching is performed for sub-domain relation extraction, and the extracted relation sentences are filtered through the semantic correcting gate. Some of the filtered relation sentences, defined as clear relation sentences, are chosen as to-be-tagged data for entity recognition.

DSFF is designed to extract entities in sentences with less work of tagging. A domain semantic feature fusion method is proposed, which learns the dependency tree construction features. Different sub-domains have similar syntactic characteristics, and it is reflected by the dependency trees. To realize syntactic feature fusion, a conditional random field (CRF) model is integrated with the dependency tree by taking the syntactic components as properties of word sequences. By this way, the CRF model could learn the domain semantic patterns. On this basis, the predicted entities are re-completed (introduced in Section IV) through an entity correcting gate.
4. SUB-DOMAIN RELATION DISCOVERY MODEL

This section introduces the Sub-Domain Relation Discovery Model (SRD Model) from three parts, sub-domain clustering part, sub-domain relationship clustering part and sub-domain relation extraction part.

4.1 Feedback Based Sub-Domain Clustering

The Feedback Based Sub-Domain Clustering aims to divide the multi-dimension featured domain into subdomains according to the industries. It uses the $tf - idf$ (Zhou X., 2020) similarity cosine distance as the sub-domain distances for clustering.

In order to analyze the similarity between the documents in the multi-dimension featured domain for subdomain clustering, they are transformed into vectors through $tf - idf$ values of Chinese word segmentation, which represents the sub-domain features of documents. Let the multi-dimension featured domain documents be:

$$D_w = \{d_1, d_2, \ldots, d_{N_{dw}}\}$$  \hspace{1cm} (1)

$d_1$ represents each document in the domain. Then process Chinese word segmentation on each document $d_1$ and remove the stop-words from it, the Chinese word group sequence of $d_1$ is obtained, supposed as:

Figure 2. Schematic diagram of the overall framework
\[ WG_{d_i} = \{ w_{1}, w_{2}, \ldots , w_{n} \} \]  

(2)

\( W_{d_i} \) represents each Chinese word group in the word sequence of \( d_i \). The set that contains all the appeared word groups is supposed as \( W_{\text{domain}} \), where \( W_{\text{domain}} \) can be expressed as:

\[ W_{\text{domain}} = \{ g_1, g_2, \ldots , g_{T_{\text{num}}} \} \]  

(3)

\( g_i \) represents each appeared Chinese word group in the multi-dimension featured domain. \( T_{\text{num}} \) represents the type number of appeared Chinese word groups. To represent the semantic features of one document through word frequency, the \( tf \) value of word group \( g_j \) in document \( d_i \) can be expressed as:

\[ tf_{ij} = \frac{TF(d_i, g_j)}{\sum_{k=1}^{T_{\text{num}}} TF(d_i, g_k)} \]  

(4)

\( TF(d_i, g_j) \) represents the times that a word group \( g_j \) appears in a document \( d_i \). The \( idf \) value is to represent the weight of each word. The \( idf \) value of word group \( g_j \) can be expressed as:

\[ idf_j = \log \left( \frac{T_{\text{num}}}{DF(g_j)} \right) \]  

(5)

\( DF(g_j) \) represents the number of documents where a word group \( g_j \) appears. The \( tf - idf \) value of word group \( g_j \) in document \( d_i \) denote as:

\[ tf - idf_{ij} = tf_{ij} \times idf_j \]  

(6)

In order to process sub-domain clustering, a domain semantic feature matrix is needed, where each row represents the semantic feature of one document. And the \( tf - idf \) matrix of the multi-dimension featured domain is the matrix, supposed as:

\[
M^{tf-idf} = \begin{pmatrix}
    tf - idf_{i1} & tf - idf_{i2} & \ldots & tf - idf_{iT_{\text{num}}} \\
    tf - idf_{i1} & tf - idf_{i2} & \ldots & tf - idf_{iT_{\text{num}}} \\
    \vdots & \vdots & \ddots & \vdots \\
    tf - idf_{iN_{doc}} & tf - idf_{iN_{doc}} & \ldots & tf - idf_{iN_{doc}T_{\text{num}}}
\end{pmatrix}
\]  

(7)

For sub-domain clustering, the \( tf - idf \) vector matrix needs to be turned into a distance matrix, which can represent the similarity of each two documents.
\[
\text{COS}\left(d_i, d_j\right) = \frac{M_{i}^{tf-idf} \cdot M_{j}^{tf-idf}}{\left|M_{i}^{tf-idf}\right| \times \left|M_{j}^{tf-idf}\right|} \tag{8}
\]

\(M_{i}^{tf-idf}\) and \(M_{j}^{tf-idf}\) represent the \(tf-idf\) vector of \(d_i\) and \(d_j\). The cosine distance matrix between domains is obtained, supposed as \(M_{\text{dis}}^{tf-idf}\).

\[
M_{\text{dis}}^{tf-idf} = \left(1 - \text{COS}\left(d_i, d_j\right)\right) \tag{9}
\]

Then hierarchical clustering is processed on \(M_{\text{dis}}^{tf-idf}\) and a dendrogram is generated. The dendrogram shows the clustering process of the domain, and the critical distance will be used to obtain the sub-domain clusters from the dendrogram with a feedback mechanism. The root node of the dendrogram is supposed as \(DG^{tf-idf}\).

\[
DG^{tf-idf} = \text{HiCluster}\left(M_{\text{dis}}^{tf-idf}\right) \tag{10}
\]

\text{HiCluster} is shown in Algorithm 1. The details of the dendrogram are introduced as follows. Suppose \(N_0\) as the root node of the dendrogram and suppose the child nodes of \(N_i\) as \(N_{2i+1}\) and \(N_{2i+2}\). Each leaf node \(N_{\text{leaf}}\) of the dendrogram corresponds to each document in the multi-dimension featured domain \(d_i\). Each non-leaf node \(N_{\text{nleaf}}\) represents a cluster, which is made up of \(N_{\text{leaf}}\) rooted sub-tree’s leaf nodes. Besides, each non-leaf node also represents one merging process between two closest clusters (two child nodes). Therefore, each node \(N_i\) in the dendrogram has a distance property which represents the distance of the two merged clusters, supposed as \(\text{Dis}\left(N_i\right)\) \(\left(\text{Dis}\left(N_{\text{leaf}}\right) = 0\right)\). The list of total nodes in the \(N_i\) rooted sub-tree is supposed as \(\text{NodeList}\left(N_i\right)\), and the list of total leaf nodes in the sub-tree is supposed as \(\text{LeafNodes}\left(N_i\right)\). They are used in \text{HiCluster}, algorithm 1.

The critical distance \(\text{CrD}_{\text{domain}}\) determines the clustering results from the dendrogram, nodes with lower merged distance than \(\text{CrD}_{\text{domain}}\) are viewed as one cluster. According to \(\text{CrD}_{\text{domain}}\) each node \(N_j\) that meets following conditions is a root node of a cluster (sub-domain).

\[
\text{Dis}\left(N_i\right) > \text{CrD}_{\text{domain}}
\]
\[
\text{Dis}\left(N_j\right) \leq \text{CrD}_{\text{domain}}, j \in 2i - 1, 2i + 1 \tag{11}
\]

To obtain the clusters (sub-domains) from dendrogram according to \(\text{CrD}_{\text{domain}}\), the function GGCD is designed as shown in Algorithm 2. After the sub-domains \(D_t\) are obtained, \(\text{CrD}_{\text{domain}}\) is adjusted according to the feedback validation \(\text{SDs}_{\text{Dbw}}\) of relationship clustering in \(D_t\). Then the best \(\text{CrD}_{\text{domain}}\) value is obtained and GGCD is repeated for the most reliable sub-domains, supposed as \(D_{\text{sub}}\).
\[ D_{sub} = \{\text{LeafNodes}(N)\} \]
\[ N \rightarrow GGCD\{NodeList(DG_{tf-idf}), 0, CrD_{domain}\} \]
\[ D_{sub} = \{S_1, S_2, \ldots, S_c\} \]

The details are introduced as follows. The feedback mechanism is to make the sub-domain clustering adapt to the relationship clustering. So, the \( CrD_{domain} \) is adjusted according to an evaluation metric, which can evaluate the sub-domain relation clustering. The \( s \_ Dbw \) (Jianhua, T., & Hongzhou, T., 2009) validation is the best validation to evaluate the clustering now, the lower, the better. But in this situation, it cannot directly evaluate the sub-domain relationship clustering due to the scale of sub-domains. When the sub-domain is large with many documents, the sub-domain self contains various industries very possibly, which leads to a broader word vector model space (Chen, H. et al., 2013a, 2013b, 2014). And relation words in different industries usually have greater word-vector distances, and the \( s \_ Dbw \) will be lower than when they are in different sub-domains. In response to this situation, the scale of each sub-domain cluster is considered as a negative factor of the validation. Here we use a validation, named as \( SDs \_ Dbw \), to evaluate the relationship clustering in subdomains. Suppose the number of documents of each subdomain \( S_i \) as \( DN(S_i) \). And the \( s \_ Dbw \) of \( S_i \) is supposed as \( s \_ Dbw_i \). The validation \( SDs \_ Dbw \) is defined as follows.

\[
SDs \_ Dbw = \left( \sum_{i=1}^{c} log_2 DN(S_i) \times s \_ Dbw_i \right) / c
\]

And it is determined by \( CrD_{domain} \), because \( CrD_{domain} \) determines the sub-domain clustering, and sub-domain relationship clustering is dependent on the sub-domain clustering result.
SUBCLU represents the sub-domain clustering process before, and SDS represents the SDS_Dbw calculation. The lower SDS_Dbw is, the better sub-domain relationship clustering is, and the better sub-domain clustering is. So CrD_doamin takes the value which makes the SDS_Dbw value the lowest.

\[ CrD_{doamin} = \arg\min_{CrD} SDS\left(\text{SUBCLU}(CrD)\right) \quad (16) \]

4.2 Sub-Domain Relationship Clustering

The Sub-Domain Relationship Clustering targets on the data in the same sub-domain \( S_i \) in \( D_{sub} \). The details are presented as follows.

To obtain the operable data from the documents, the sub-domain data \( S_i \) is split into sentences, supposed as \( S_{data} \).

\[ S_{data} = \{ s_1, s_2, \ldots, s_n \} \quad (17) \]

In order to obtain the relation words of sentences, for each sentence \( S_i \) in \( S_{data} \), dependency parsing is processed, and its dependency tree is obtained, supposed as \( T_i \).

\[ T_i = f_{SA}\left(s_i\right) \quad (18) \]

\( f_{SA}\left(s_i\right) \) represents the dependency parsing on \( S_i \).

Node ROOT and Node \( ADV(RO\text{OT}) \) of each tree \( T_i \) are extracted as relation words of \( S_i \), because most of the relations are expressed with restrictive words, mostly as Node Root or Node \( ADV(RO\text{OT}) \). This extraction function is defined as \( Rule_r\left(T_i\right) \). Then we get relation word set \( R_{base} \).

\[ R_{base} = \{ Rule_r\left(T_1\right), Rule_r\left(T_2\right), \ldots, Rule_r\left(T_n\right) \} \quad (19) \]

\[ R_{base} = \{ r_1, r_2, \ldots, r_m \} \quad (20) \]

\( r_i, i \in \{1,2,\ldots,m\} \) represents each relation words in \( R_{base} \). The relation words will be transformed into vectors for relationship clustering, through word embedding. For word embedding, the nodes of \( T_i \) are taken in order, supposed as \( Nodes\left(T_i\right) \), as the word group sequence, defined as \( W_i \). And the word sequence of the source data is supposed as \( W_{total} \).

\[ W_{total} = \{ Nodes\left(T_1\right), Nodes\left(T_2\right), \ldots, Nodes\left(T_n\right) \} \quad (21) \]

Then word embedding is processed, using skip-gram model and training with \( W_{total} \). The model is supposed as \( M_{total} \).

\[ M_{total} = T_{skip-gram}\left(W_{total}\right) \quad (22) \]
T_{skip-gram} represents the training process with skip-gram model. In the process of word embedding, suppose one-hot encoding length as \( V \), the length of embedding level as \( N \), then the word-vector of each word group \( r_i \) in relation word set \( R_{base} \) is obtained, supposed as \( v_i \).

\[
v_i = \text{Index}_{r_i} \left( M^{W_{v,x}}_{\text{total}} \right) \tag{23}
\]

\( M^{W_{v,x}}_{\text{total}} \) represents the embedding level of the model, which contains the representations of embedded words through the unsupervised learning of the skip-gram model. \( \text{Index}_{r_i} \) represents the word-vector corresponding to \( r_i \). Suppose the relation word-vector set as \( V_r \).

\[
V_r = \left\{ v_1, v_2, \ldots, v_m \right\} \tag{24}
\]

Then it comes to sub-domain relationship clustering for the word vectors that have the relation features. Hierarchical clustering is processed on relation word-vector set \( V_r \) according to cosine distances. The dendrogram of the hierarchical clustering is obtained, and the root node of which is supposed as \( D G_{\text{ROOT}} \). The cosine distance matrix of \( V_r \) can be obtained by the same way as in sub-domain clustering model, supposed as \( M_{V_r}^{\text{dis}} \).

\[
M_{V_r}^{\text{dis}} = \left\{ 1 - \cos \left( v_i, v_j \right) \right\} \tag{25}
\]

\[
D G_{\text{ROOT}} = \text{HiCluster} \left( M_{V_r}^{\text{dis}} \right) \tag{26}
\]

The symbols above are introduced before in sub-section A. The dendrogram has been introduced in sub-section A as well and the critical distance of the hierarchical clustering here is supposed as \( CrD \). The ideal critical distance is a little higher than the densest cluster-merged distance ranges, so the relation words can be clustered in large density groups. But in some sub-domains, which have a small number of relation words, the above critical distance makes confusion. Different types of relation words will be clustered in the same group, because the word-vectors usually have great distance close to the maximum. In this situation, \( CrD \) takes the following value.

\[
CrD = \text{MIN} \left\{ 1.1 \times \sum_{i=0}^{n} \frac{\text{Dis}(N_i)}{n}, 0.9 \times \text{Dis}(N_0) \right\} \tag{27}
\]

\( \text{Dis}(N_0) \) represents the distance of merging clusters of \( N_0 \left( \text{ROOT Node} \right) \). The clustering result is supposed as \( C_r \).

\[
C_r = \left\{ \text{LeafNodes} \left( N \right) \right\} \tag{28}
\]

\[
N \rightarrow \text{GGCD} \left( \text{NodeList} \left( D G_{\text{ROOT}} \right), 0, CrD \right) \tag{29}
\]

\[
C_r = \left\{ c_1, c_2, \ldots, c_k \right\} \tag{29}
\]
represents each relation word cluster. Then the trigger word set of sub-domain $S_i$ is obtained, supposed as $R_{\text{trigger}}$.

$$R_{\text{trigger}} = \{w | w \in c, c \in C_r\}$$  \hspace{1cm} (30)

Figure 3 shows the process of Sub-Domain relationship Clustering.

**Algorithm 1** Hierarchical Clustering on Distance Matrix

**Input:** Distance Matrix of Certain Vector Set  
**Output:** The Root Node of The Clustering Dendrogram  

function HiCluster($M$)

$n = M.\text{size}$  

$\text{Node}(N_1, N_2)$: return a node with child nodes $N_1$ and $N_2$  

$\text{NodeList} = \{\text{Node}(\text{null}, \text{null}), \text{Node}(\text{null}, \text{null}), \ldots, \text{Node}(\text{null}, \text{null})\}$  

while $\text{NodeList.\text{size}} > 1$

search min value $d_{k,l}$ in $M$

$\text{NewNode} = \text{Node}(N_k, N_l)$  

$\text{NewNode.\text{dis}} = d_{k,l}$  

remove $N_k$ and $N_l$ from $\text{NodeList}$  

$\text{NodeList} \rightarrow \text{NewNode}$

update $M$

end while

return $\text{NewNode}$

end function

**4.3 Sub-Domain Relation Extraction**

For each $T_i$ of $s_i$ in $S_{\text{data}}$, match node $\text{ROOT}(T_i)$ and node $\text{ADV}\left(\text{ROOT}(T_i)\right)$ with trigger words in $R_{\text{trigger}}$, supposed as $f_{\text{match}}^i$.

$$f_{\text{match}}^i = \text{ROOT}(T_i) \in R_{\text{trigger}} \land \text{ADV}(\text{ROOT}(T_i)) \in R_{\text{trigger}}$$  \hspace{1cm} (31)

If $f_{\text{match}}^i$ is true, then push sentence $s_i$ and tree $T_i$ in a buffer $B_{\text{imp}}$. Then sentences in $B_{\text{imp}}$ will be filtered, based on the constraints of dependency tree construction, supposed as $\text{FLT}(s_i)$. For example, a relation instance must have a specific right entity, so the dependency tree must have node $\text{VOB}(\text{ROOT})$. In this case, $\text{VOB}(\text{ROOT})$ cannot be a pronoun or some non-clear expression. With a list of non-clear right entity word groups, the non-specific relations will be filtered out. Then the clear relation sentences of sub-domain $S_i$ are obtained, supposed as $RI_{s_i}$.

$$RI_{s_i} = \{s_i | s_i \in S_i, f_{\text{match}}^i = \text{True}, \text{FLT}(s_i) = \text{True}\}$$  \hspace{1cm} (32)

Figure 4 shows the process of Sub-Domain Relation Extraction.

**Algorithm 2** Get Clusters from Dendrogram of Hierarchical Clustering with Critical Distance
Input: $N$ is the node set of the dendrogram, $\text{index}$ is the node index, $\text{CrD}$ is the critical distance
Output: The root node set of each cluster

function GetClustersFromDen($N$, $\text{index}$, $\text{CrD}$)

Set$_{\text{clu}} = \{\}$

if $\text{Dis}(N_{\text{index}}) < \text{CrD}$ then return Set$_{\text{clu}}$
end if

LTag = False
RTag = False

if $\text{Dis}(N_{2 \times \text{index} + 1}) < \text{CrD}$ then
    Set$_{\text{clu}} \leftarrow N_{2 \times \text{index} + 1}$
    LTag = True
end if

if $\text{Dis}(N_{2 \times \text{index} + 2}) < \text{CrD}$ then
    Set$_{\text{clu}} \leftarrow N_{2 \times \text{index} + 2}$
    RTag = True
end if

if LTag & RTag then return Set$_{\text{clu}}$
end if

if LTag & not RTag then return Set$_{\text{clu}} \cup \text{GetClustersFromDen}(N, 2 \times \text{index} + 2, \text{CrD})$
end if

if not LTag & RTag then return Set$_{\text{clu}} \cup \text{GetClustersFromDen}(N, 2 \times \text{index} + 1, \text{CrD})$
end if
return GetClustersFromDen($N, 2^{index+1}, CrD$) \hat{=} GetClustersFromDen($N, 2^{index + 2}, CrD$)
end function

5. DOMAIN SEMANTIC FEATURES FUSION MODEL

The Domain Semantic Features Fusion Model (DSFF Model) aims to extract the entities in the clear relation sentences, tagging as little as possible. A dependency parsing construction fusion model is designed to learn the semantic features of the multi-dimension featured domain. Here we use CRF model for semantic features fusion. The input word sequence is the node sequence in the dependency tree $T_i$. And the syntactic compositions are taken as the property of each Chinese word group, such as $ROOT$, $ATT$, $SBV$ and so on. The syntactic composition sequence is supposed as $SC$.

First, choose some clear relation sentences in $RI_{S_i}$, supposed as $S$.

$$S = \{s_1, s_2, \ldots, s_m\}$$ (33)

For $T_i$ of $s_i$ in $S$, tag the entity nodes, supposed as $Ne(T_i)$. The tagging process is supposed as $f_{tag}(s_i)$. Then the tag sequence of $s_i$ is obtained, supposed as $tg_i$.

Figure 5.
Schematic diagram of domain semantic features fusion model
Then process CRF training and prediction to get the entities.

\[
\begin{align*}
    t_g &= f_{tag}(s_i) \\
    S_t &= \{ t_g, t_g_2, \ldots, t_g_m \}
\end{align*}
\]  

\[
\begin{align*}
    Model_{CRF} &= \text{Train}(S, SC, S_t) \quad (36) \\
    \text{PreE} &= \text{SeqInfer}(Model_{CRF}, RI_s) \quad (37)
\end{align*}
\]

\( \text{Train}(segs) \) represents the CRF training on the input sequences \( segs \). \( \text{SeqInfer} \) represents the sequence inference. The prediction entities \( \text{PreE} \) are composition completed through the entity correcting gate. It is designed with attention to the syntactic component sequence. Some clear relation sentences have separated parts in one entity, such as \{管理, 系统, 和, 支撑, 环境, 是, 数据, 存储, \', 管理, 和, 运行, 维护, 的, 软硬件, 及, 网络, 条件\} \{management, system, and, support, Environment, is, data, storage, \', Management, and, executing, maintenance, hardware and software, and, network, conditions\}, the entities 数据管理 (data management), 数据运行 (data executing) and 数据维护 (data maintenance) are separated. In response to this, the composition completing gate connects the separated parts, through constraints on the syntactic components of dependency trees. Figure 5 shows the process of Domain Semantic Features Fusion Model.

6. EXPERIMENTS AND RESULTS

6.1 Dataset

The proposed method is evaluated on publicly available documents of NSNR, which can be obtained from the official website of Ministry of Natural Resources of the people’s Republic of China. We used 172 files of them, including national standards for land, geology and minerals, national marine standards, and national surveying and mapping standards.

6.2 Experiment Configuration

All the experiments are implemented using Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz (12 CPUs) in Windows 10 OS, 16GB RAM.

6.3 Sub-Domain Clustering

In order to show the multi-dimension features of the NSNR domain and for sub-domain clustering, \( tf – idf \) vector hierarchical clustering are processed on the standard documents. The dendrogram of it is shown in Figure 6. It can be noticed that most documents have a great cosine distance in

<table>
<thead>
<tr>
<th>Group Index</th>
<th>Relation Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td>采用 用于 详细 提供 使用 包含 包括 注 表示 说明 应 进行 保存 引用</td>
</tr>
<tr>
<td>Group2</td>
<td>是 定义 为 描述</td>
</tr>
<tr>
<td>Group3</td>
<td>表达</td>
</tr>
</tbody>
</table>

A sample of relation clustering of certain sub-domain in situation 2
The feedback mechanism on sub-domain clustering $CrD_{domain}$ is shown in Figure 7. The $CrD - SDs_{Dbw}$ curve first shows a downward trend and then an upward trend, which accords with the actual case. It can be told that the best $CrD_{domain}$ is 0.81, which is in line with expectations, a little greater than the densest cluster merged distances. The sub-domain clusters in this case have the best relationship clustering effect. It can be also told that when the $CrD_{domain}$ comes far from the extreme value 0.81, the $SDs_{Dbw}$ value tends to rise up. Here we discuss the extreme situation of $CrD_{domain}$ and its influence to the proposed metric $SDs_{Dbw}$. When $CrD_{domain}$ comes very high (close to 1), all the documents will be gathered in one same sub-domain. The weight of the penalty item DN (sub-domain document numbers) in equation (14) will be extremely large (the number of total documents) and the $SDs_{Dbw}$ will become bad (high) as well. On the other hand, when $CrD_{domain}$ comes extremely low (close to 0), all the sub-domains self contain only one documents, and the metric of relation clustering must be extremely bad. Here the metric $s_{Dbw}$ of each sub-domain in equation (14) becomes high and thus the proposed metric $SDs_{Dbw}$ will become bad (high) as well. Through the discussion on extreme situations above, it can be reflected that the proposed metric $SDs_{Dbw}$ which evaluates the relation clustering of multi-dimensional domain are reasonable and reliable.

The proposed metric $SDs_{Dbw}$ is more than reasonable, but also flexible. The utilized clustering metric $s_{Dbw}$ can be replaced by any other reasonable metric, such as AMI, ARI, V-measure, et al., as long as it can evaluate the relation discovery of each sub-domain, including but not limited to clustering method metrics. The kernel of the metric $SDs_{Dbw}$ is the penalty of its expression, the scale of each sub-domain, when the relation discovery metric comes higher to be better which is different from $s_{Dbw}$, the penalty item can be put into the fraction in consistent with the relation discovery metric. This pattern can be utilized in any multi-dimensional task when the scale of each sub-clusters make influence in the next options.

6.4 Relationship Clustering

There are usually two types of situations in the hierarchical clustering on sub-domain data.

Situ. 1: One is that the sub-domain has obviously various relation types with many relation words. The cluster merged distances are mostly low and individually high. Because distances between relation word-vectors in the same type are close, and distances between different types are high.

Situ. 2: The other situation is that the sub-domain which contains only a few documents, even only one or two. It has some relation types, and each one has only a few relations. Therefore, the distances among the relation word-vectors are mostly large, even close to the largest one in the dendrogram.

The two situations are shown at Figure 8. $CrD$ in Situ. 1 should be between mostly low distances and individually high distances, while $CrD$ in Situ. 2 should be lower than the highest distance. In response to this, the $CrD$ of the relation hierarchical clustering takes the minimum of two designed validations.
Figure 6.
Hierarchical clustering dendrogram on the tf-idf vector of the planning domain documents

![Hierarchical Clustering Dendrogram](image)

Figure 7.
Feedback mechanism on determining the best $CrD_{domain}$

![Feedback Mechanism on Sub-Domain Clustering](image)
CrD = MIN \left( 1.1 \times \sum_{i=0}^{n} \text{Dis}\left( N_i \right) + 0.9 \times \text{Dis}\left( N_0 \right) \right) \quad (38)

Usually the Situ. 2 is more difficult to deal with. There is a relation clustering sample of a sub-domain in Situ. 2 shown in Table I, which takes CrD above. It can be told that the relation words, which means ‘use’ and ‘consist’, are clustered together and the relation words, which means ‘define’, are clustered together. This is close to manual classification.

6.5 Entity Recognition on Domain Semantic Features Fusion Model

To test the effect of the semantic fusion entity recognition model, we tag some clear relation sentences in a sub-domain. Clear relation sentences in one document are taken as training data, those in other four documents are testing data. The entity counting recognition task is processed for the performance contrast between the DSFF model and the traditional CRF model. The labeled entities in the training

<table>
<thead>
<tr>
<th>Document</th>
<th>DSFF Model</th>
<th>Traditional CRF Model</th>
<th>POS Tags Fusion Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPGIQ</td>
<td>0.78</td>
<td>0.49</td>
<td>0.60</td>
</tr>
<tr>
<td>BGID</td>
<td>0.87</td>
<td>0.55</td>
<td>0.67</td>
</tr>
<tr>
<td>ETIC</td>
<td>0.86</td>
<td>0.55</td>
<td>0.67</td>
</tr>
<tr>
<td>IGDE</td>
<td>0.79</td>
<td>0.43</td>
<td>0.56</td>
</tr>
<tr>
<td>EPGIQ</td>
<td>0.77</td>
<td>0.42</td>
<td>0.54</td>
</tr>
<tr>
<td>BGID</td>
<td>0.91</td>
<td>0.55</td>
<td>0.69</td>
</tr>
<tr>
<td>ETIC</td>
<td>0.87</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>IGDE</td>
<td>0.80</td>
<td>0.42</td>
<td>0.55</td>
</tr>
</tbody>
</table>
documents will be fed to the models and the predicted entities in testing data will be counted with both correct ones and error ones. Precision, Recall and F-Measure are used for comparison. The sequence of about 4000 words are taken as training data, which is split into 10 equal parts in order. Different scales of training data are taken randomly form the 10 split parts, fed into DSFF model separately in 5, 6 and 7 parts.

\[
P = \frac{\text{Truepositives}}{\text{Truepositives} + \text{Falsepositives}}
\]

\[
R = \frac{\text{Truepositives}}{\text{Truepositives} + \text{Falsenegatives}}
\]

\[
F = \frac{2 \times P \times R}{P + R}
\]

There is a comparison among semantic feature fusion model, word group sequence without properties model and POS tags fusion model in Table II. We compare the performances in two tasks, one is on the entity types which focuses on the recognized types of entities and ignores the appeared times of each entity. The other is on the entity numbers which counts the appeared times of each entity. It can be told that the semantic feature fusion model has more advantage than the other 2 models in entity type recognition, especially for recall validation, which confirms that the semantic feature fusion model has stronger ability in extract new syntactic-construction similar entities. The POS fusion model sometimes has higher recall on the total numbers of extracted entities but always has a lot lower precision than the other two models, which indicates that the single POS fusion model may cause confusion in recognizing the subject entity in a relation instance, because it cannot learn the syntactic features.

To test DSFF model’s entity recognition ability with less training data, we use part of the tagged data CDUG as training data of domain semantic feature fusion model (DSFF). The parts of the tagged data are respectively 50%, 60% and 70%, and the performances between DSFF model and traditional CRF model are shown in Table III. It can be told that with 60% training data, DSFF model has better performances in recall and F-measure than traditional CRF model on all tested documents, which confirms that DSFF model has a stronger ability in recognizing the entities through syntactic construction features. Compared to the traditional entity recognition model, it can reach the same or better performance with less training data.

### 7. CONTRIBUTIONS

The contributions of this paper are as follows:

<table>
<thead>
<tr>
<th>Document</th>
<th>DSFF Model</th>
<th>Traditional CRF Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50% training data</td>
<td>60% training data</td>
<td>70% training data</td>
</tr>
<tr>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>EPGIQ</td>
<td>0.68</td>
<td>0.38</td>
</tr>
<tr>
<td>BGID</td>
<td>0.81</td>
<td>0.43</td>
</tr>
<tr>
<td>ETIC</td>
<td>0.89</td>
<td>0.47</td>
</tr>
<tr>
<td>IGDE</td>
<td>0.65</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 3. Entity Recognition Performance with Different Amount of Tagged Data
1. This paper extracts structural knowledge from National Standards for Natural Resources (NSNR), which facilitates the revision and execution of NSNR to help make better utility of China’s natural resources.

2. A multi-domain splitting model with feedback mechanism is proposed for relation extraction from NSNR, which could automatically split the multi-domain texts into sub-domain texts with little demand of expert efforts.

3. A new evaluation metric based on sub-domain relationship clustering is designed to better split multi-domain texts, which is experimentally demonstrated to be effective.

4. An entity recognition method with semantic features fusion is designed to learn the multi-dimensional semantic features, which reaches good performance with less training data and greatly reduces the workload of manual tagging.

8. CONCLUSION

This paper proposes a knowledge extraction method for multi-domain texts to extract structural knowledge from National Standards for Natural Resources (NSNR). The proposed method addresses some challenges of knowledge extraction from NSNR: 1) The domain is made up of different sub-domains. 2) Each sub-domain is interdisciplinary. 3) Different sub-domains are partly similar. 4) There lacks tagged corpus. Concretely, a multi-dimension splitting method with feedback mechanism is proposed for knowledge discovery of the multi-domain texts. For the feedback mechanism, $SDs_{Dbw}$ validation is designed with the consideration of wrong merging risks to evaluate sub-domain relation clustering. In sub-domain relation clustering, a CrD valuing method is designed for automatic relation discovery. For entity recognition, we use a semantic feature fusion CRF model through dependency parsing feature fusion. Experimental results demonstrate that the proposed method has good effect on relation discovery and entity recognition of NSNR. However, the method still has certain limitation. When the data covered by the multi-dimensional domain is too extensive, certain entity features of some sub-domains could be lost. In future work, some entity frequency-related indicators could be explored and integrated to break this limitation.

ACKNOWLEDGMENT

Thanks are due to Lyuzhou Chen for valuable discussion.

COMPETING INTERESTS

The authors of this publication declare there are no competing interests.

FUNDING AGENCY

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the author(s) of the article.
REFERENCES


**ENDNOTES**

1. The National Standards for Natural Resources contain standards for development of marine resources, exploitation of ore resources, urban planning, surveying, and mapping (Chen, H. *et al.*., 2010b, 2010c) etc. which are important for the country’s natural resource utilization.
2. For example, the standards of geographic information are described from system construction specification, quality evaluation standard, data product specification, etc., which are related to Geography Science and Computer Science.
3. Such as standard relevance analysis, the contradiction analysis between standards, the merge and modification of standard.
4. [http://gi.mnr.gov.cn/201904/P020190428326425854439.xlsx](http://gi.mnr.gov.cn/201904/P020190428326425854439.xlsx)
Taiyu Ban received the bachelor’s degree from USTC, Hefei, China, in 2020. He is currently pursuing the Master’s degree with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include machine learning and optimization theory.

Xiangyu Wang received the B.Sc. degree from Donghua University, Shanghai, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Data Science, University of Science and Technology of China, Hefei, China. His research interests include machine learning and optimization theory.

Xin Wang received the bachelor’s degree from USTC, Hefei, China, in 2019. He is currently pursuing the Master’s degree with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include machine learning and optimization theory.

Jiarun Zhu received the bachelor’s degree from USTC, Hefei, China, in 2019. He is currently pursuing the Master’s degree with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include machine learning and optimization theory.

Lvzhou Chen received the bachelor’s degree from USTC, Hefei, China, in 2020. He is currently pursuing the Master’s degree with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include machine learning and optimization theory.

Yizhan Fan received the bachelor’s degree from USTC, Hefei, China, in 2020. He is currently pursuing the Master’s degree with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include machine learning and optimization theory.