International Journal of Civil and Structural Engineering – IJCSE

Volume 6 : Issue 1- [ISSN : 2372-3971] - Publication Date: 10 May, 2019

Rewriting of Design Code With Controlled Natural Language to Assist Information Extraction

[Bing Wu[1], Yuanbin Song[2], Ruoxin Xiong[2], Fulin Li[2]]

***Abstract*—Regulatory compliance checking is crucial for power grid designs with respect to their safe operation. Traditional design checking depends greatly on the knowledge and experiences of human expert, and accordingly the manual checking is often tedious, time consuming and error prone. Recently, the application of Natural Language Processing (NLP) approach to extract information from design codes written in English leads to promising results. Unfortunately, the robustness of NLP-based information extraction approach should be furthered in order to analyze the textual codes written in Chinese. In this regard, GIM-CNL is developed as a controlled language for rewriting Chinese design code for power grid infrastructures. This study explores the difficulties in understanding design code written in Chinese. The framework of the GIM-CNL and the associated ontology of power grid engineering is also developed. Finally, an example of information extraction from the design codes is studied to demonstrate the application of the proposed tool.**

***Keywords*— Natural language processing, Controlled natural language, design code, information extraction, and ontology**

I. Introduction

The design of a power grid engineering project is constrained by a number of design and construction regulations, such as Code for Design of High Voltage Electrical Installation (GB50060-2008), Design Codes for Electrical Power Engineering (DL 503-1992), Code for Design of Cables of Electric Work (GB 50217-1994) and Technical code for Construction Survey in Electric Power Engineering (DL/T 5445-2010). In addition, the power industry should also complies with a number of design code enforced by other industries. Besides the large amount of textual regulatory documents for design and construction, the inconsistency of their semantics representation and writing style further exacerbate the difficulties of extracting design rules since manual extraction of design checking rules is often costly, time consuming, and error prone [1]. On the other hand, a power grid project, failing to comply with regulatory design and construction codes, might incur high expenditure of construction and operation.

In the last decade, many academic efforts have been paid to explore the application of automated tools in order to release the human experts from the time consuming works of compliance checking [2,3]. Meanwhile, a number of tools have been developed for automatically checking designs against design and construction codes, like the CORENET by the Singapore Ministry of National Development [4], SMART by the International Code Council [5], EPLAN/BIM provided by FIATECH [6], Solibri Model Checker [7], and Avolve [8].

Although these compliance checking tools can save designers or engineers a lot of efforts from manually checking the design model against regulatory rules, many of these regulatory rules were still manually encoded or scripted into the computer systems [9]. Subsequently, Zhang and El-Gohary [10] developed a method for automatic extraction checking rules from design codes in English with assistance of Natural Language Processing (NLP) technology, but the developed approaches may not be adaptive for codes written in Chinese because of the different organization and structure between those two languages. This paper proposes a framework to facilitate the information extraction from design code rewritten in controlled Chinese natural language with the help of NLP tools. In particular, the framework of controlled natural language GIM-CNL and the ontology of power grid engineering has been developed to improve code understanding for better results of rule extraction.

II. Application of NLP for Information Extraction

NLP is the approach concerning how to use computer systems to process and analyze non-structural documents written in natural language. Automatic information extraction, a subfield of natural language understanding, is one of the typical challenges in natural language processing. Early research focused on rule-based tools where many computer programs were developed with a set of processing rules manually coded, for instance, heuristic rules for stemming. However, such an approach is not in general robust to the variation of textual representation. Later, statistical means was explored by using statistical learning approaches to automatically derive rules through the analysis of large corpora. Recently, NLP tools based on deep learning can

Bing Wu[1]
Economic and Technological Research Institute, State Grid Zhejiang Electric Power Co(Ltd.)
China

Yuanbin Song[2]
School of Naval Architecture, Ocean and Civil Engineering
Shanghai Jiaotong University
China

Ruoxin Xiong[2]
School of Naval Architecture, Ocean and Civil Engineering
Shanghai Jiaotong University
China

Fulin Li[2]
School of Naval Architecture, Ocean and Civil Engineering
Shanghai Jiaotong University
China

achieve state-of-art performance in complex tasks, like machine translation and question answering.

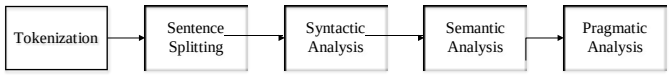

Figure 1. Example of a figure caption.

In general, the typical pipeline of NLP often consists of a sequence of function modules as shown in Figure 1.

(1) Tokenization segments a sentence into a sequence of parts, called tokens, for further processing.

(2) Sentence splitting is used to recognize and split sentences, depending on such punctuation marks as periods, exclamation marks, and question marks.

(3) Syntactic analysis derives the structure or grammar of a single sentence.

(4) Semantic analysis derives the meaning of a phrase or a sentence.

(5) Pragmatics analysis concerns how to reference the meaning of a sentence to its purpose of communication.

Figure 2. Segmentation and POS of Example Chinese Sentence.

Figure 3. Phrase structure analysis of example sentence for clearance constraint.

Since there is no space separator between Chinese characters as English sentences, this study employs HanLP [11] for processing the Chinese sentence. In our tests of processing design codes, HanLP can better segment a sentence into a sequence of words than other NLP tools. Figure 2 illustrates the segmentation of the example code sentence, being tagged by part of speech (POS), while Figure 3 shows the tree structure of the phrases of the same sentence as that in Figure 2. The English translation of the Chinese sentence in Figure 2 is that the clearance distance between 110kv electrified components and non-charged components outside the building should not be less than 1000 millimeters. In addition, the meaning of the POS labels and the relationships of phase used in Figures 2 and 3 are explained in Table 1.

TABLE I. POS AND RELATIONSHIPS BETWEEN PHRASES TAGS

Abbreviation	Description	Abbreviation	Description
c	conjunction	ADV	adverbial
d	adverb	ATT	attribute
m	number	COO	coordinate
n	general noun	HED	head
nd	noun for direction	LAD	left adjunct
p	preposition	POB	preposition-object
u	auxiliary	RAD	right adjunct
v	verb	SBV	subject-verb
wp	punctuation	VOB	verb-object

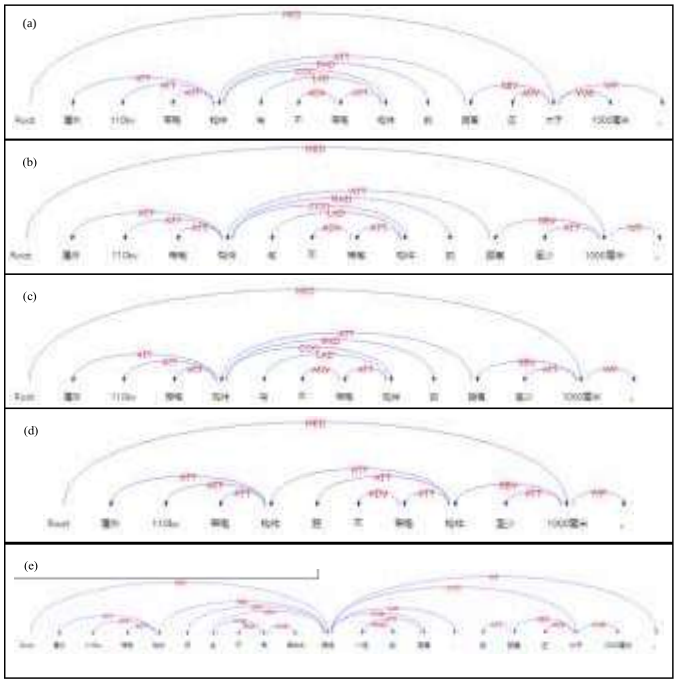

Figure 4. Phrase structure of five sentences with the same meaning.

Similarly, due to the flexible organization of Chinese Language, there are a number of alternatives to paraphrase the same design constraint in the regulatory documents. For instance, the code item of clearance distance studied in Figures 2 and 3 can be rephrased by at least another 5 sentences with their phrase structures illustrated in (a) to (e) of Figure 4. In all these five sentences, the location modifier “屋外” (outside the building) is ambiguous. Although human engineers can understand it is adverbial for the sentence, the computer system may misunderstand it as attribute for the first noun “构件” (building component). In contrast, “在建筑物外” stated in the example sentence in Figure 2 is explicitly adverbial. “屋” is also ambiguous in Chinese that either building or room can be explained relying on the context. Furthermore, the phrases “至少” in Figures 4(b) to 4(d) have the same meaning as “应大于” in Figure 4(e). Again, the sentence studied in Figures 4(e) is comparatively complex for computer to analyze, and frequently error prone. In this regard, the controlled natural language GIM-CNL, short for Grid Information Modeling Controlled Natural Language, is proposed to reduce the ambiguity and misunderstanding of the design and construction code.

III. Controlled Natural Language

A. GIM-CNL

Natural languages are probably the most expressive approach to represent design and construction code since they are convenient for engineers to use and understand. Unfortunately, it is this expressive capability that makes natural languages extremely tough for a computer to extract checking rules from existing design codes since a lot of ambiguity of computer understanding exists and background knowledge is usually not stated in the regulatory documents. Therefore, the controlled natural language and ontology are specifically developed to resolve the aforementioned two problems, respectively.

Various types of programming languages and Domain Specific Languages (DSLs) have been suggested to represent field knowledge because of their well-defined syntax, unambiguous semantics and automated inference capability. Nevertheless, these computer readable languages are often difficult for designers or engineers to understand and apply. Controlled Natural Language (CNL) can bridge the gap between a natural language and a computer language. A CNL is the subset of a natural language with its vocabulary and grammar have been intentionally limited and systematically devised in order to reduce both ambiguity and complexity of the corresponding full natural language.

Previous research categorized CNLs into human-oriented and machine-oriented CNLs [12]. While the former, like ASD -STE100 Simplified Technical English[13], focuses on improving the readability and understandability of technical documentation, the latter, like Semantic Web [14], attempt to improve the translatability of technical documents for better knowledge representation.

The GIM-CNL has been developed as a friendly interface for designers/engineers to rephrase the existing design and construction codes, and also as a knowledge representation schema for computers to process and understand textual codes. In detail, the design code rewritten in GIM-CNL by experts can be translated by computer into DSL rules. In this way, the compliance of power grid designs can be automatically checked against those DSL rules stored in the knowledgebase.

B. Vocabulary of GIM-CNL

The GIM-CNL vocabulary consists of the followings:

Domain Term: A domain entity represented by a common noun or a noun phrase. For example, a building is a spatial object, while non-charged parts are physical objects.

Constraint Term: denoting the engineering requirement for an building element, its attribute(s), or a relationship between building elements.

Domain Attribute: A determiner indicating such attributes of an entity in an engineering design as function, material, shape, process.

Domain Verb: restricting a relationship, an attributes or action involving one or more domain entities.

Measurement Unit: Typical examples are ampere, voltage, meter, and cubic meter.

C. Grammar of GIM-CNL

GIM-CNL grammar defines syntax and constrains structure for writing sentences of design codes. Only the words in the GIM-CNL vocabulary can be used and referenced by the grammar. The general structure of representing an engineering constraint is organized as the following:

[Substructure] Constraint Term + Modality + Domain Verb + Quantity + Measurement Unit.

The substructure can further contains low-level substructures, and in this way the general pattern of a substructure can be organized from a simple to a compound structure. The constraint term serves as the subject of the sentence, while the substructure can be used to depict the conditions and/or scopes to restrict the constraint term. Each sentence allows only one constraint term, and only one domain verb, serving as the predicate of the sentence, which can be followed by quantity and measurement unit.

For example, in the Chinese sentence “在建筑物外带电构件与不带电构件之间的净空距离应大于 1000 毫米。”, “净空 距 离 ” (clearance distance) is a constraint term serving as the subject of the sentence, while “应大于” is the predicate of the sentence, restricting the clearance distance. “1000 毫 米 ” are two words, indicating quantity and measurement unit. In addition, “在建筑物外” is a substructure to restrict spatial scope of the clearance distance, the constraint term “净空距离 ”. Meanwhile, “带电构件与不带电构件之间的” is another substructure to determine the constraint term. In this regard, the complexity of code representation is concentrated on the description of the constraint term, and therefore the ambiguity of natural language understanding is greatly reduced.

D. Domain Ontology Associated with GIM-CNL

The ontology of grid engineering knowledge is also developed and associated with the restricted vocabulary of GIM-CNL. This ontology describes the conceptual model of the power grid engineering knowledge which defines a collection of interrelated concepts that are necessary for a computer to understand the design codes. Each word or terminology in the GIM-CNL vocabulary must be associated with a concept in the power grid engineering ontology partially illustrated in Figure 5.

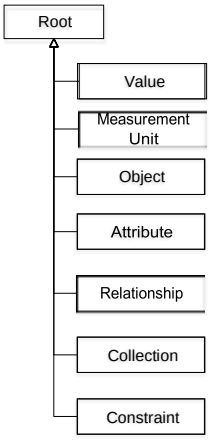

Figure 5. High Level Model of grid engineering ontology.

Figure 5 illustrates the high level model of grid engineering ontology, containing such concept as value, measurement unit, object, attribute, relationship, collection, and constraint. The ontology model of the power grid engineering is a network that interrelates these concepts with such relationships as " subtype of", "related" "relating", "define", "member of", and etc..

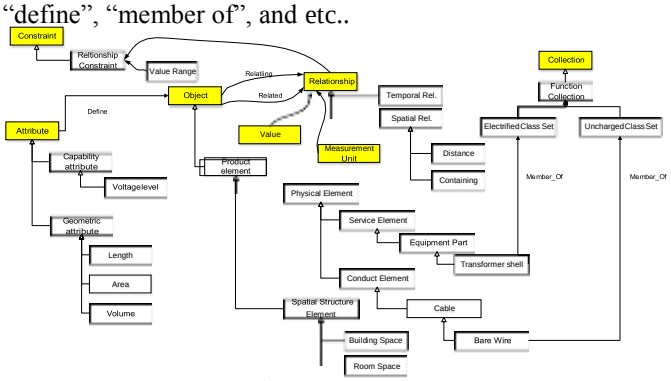

Figure 6. Partial ontology for explaining clearance distance constraints.

Figure 6 demonstrates partial ontology for explaining the clearance distance constraints that the building objects are depicted by various types of engineering attributes, and these objected are connected into a network by the relationships between objects/attributes. In this way, the grid engineering ontology can be integrated with the GIM-CNL for computers to understand design checking rules described in the design codes.

For instance, the example sentence in Figure 2 can be further evaluated, based on the POS and phrase structure analysis as shown in Figure 3, to generate a list of GIM-CNL phrases, and then their corresponding ontology concepts can be located in the GIM-CNL vocabulary, as well as their ancestor concepts (see Table II).

TABLE II. GIM-CNL PHRASES AND ONTOLOGY CONCEPT

Original Word	GIM-CNL Phrase	Ontology Concept	Ancestor Concept
在	在……外	Spatial Rel.	Relationship
外	-	-	-
建筑物	建筑物	Building Space	Object
110	110	Integer	Quantity
千伏	千伏	Electricity Unit	Unit
带电	带电构件	Electrified Class Set	Collection
构件	-	-	-
与	与	Coordinate	Relationship
不	不带电构件	Non-charged Class Set	Collection
带电	-	-	-
构件	-	-	-
的	的	-	-
净空	净空距离	Spatial Rel.	Relationship
距离	-	-	-
应	应	Requirement	Constraint
大于	大于	Comparative Rel.	Relationship
1000	1000	Integer	Quantity
毫米	毫米	Length Unit	Unit

With the predefined machine translation rules, the inference engine can translate the studied sentence into the following DSL rule:

Larger_than(" 净 空 距 离 " (Relating_Object, Related_Object), 1000, "毫米")::
WHERE(Relating_Object)
{
ATTRIBUTE("额定电压")== "110 千伏";
TYPE .IN. COLLECTION("带电");
OUTSIDE(SELF, FIND_ALL_OBJECT("建筑物").BOUNDARY);
};
WHERE(Related_Object)
{
TYPE .IN. COLLECTION("不带电");
OUTSIDE(SELF, FIND_ALL_OBJECT("建筑物").BOUNDARY);

IV. Conclusions

To resolve the ambiguity in understanding design code by computer, the controlled natural language GIM-CNL is developed. In particular, the restricted vocabulary and grammar can be used to bridge the gap between natural and the domain specific language, and therefore the GIM-CNL provides a powerful tool for rephrasing the existing design codes required for checking power grid designs. Then, the existing NLP tools can better parse the rewritten design code, and subsequently the computer system can automatically translate the CNL design codes into DSL rules. In this way, a lot of manpower traditionally spent on manual development of compliance checking rules can be saved. In the future, we may enhance the understanding capacity of the automatic

translation system to deal with more types of rules other than clearance distance.

Acknowledgment

The authors would like to acknowledge the sponsorship of Economic and Technological Research Institute, State Grid Zhejiang Electric Power Co (Ltd.) (Grant No. ZBWZ18-012-003) and the financial support of the National Natural Science Foundation of China (Grant No.71271137). Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily represent those of the Economic and Technological Research Institute, State Grid Zhejiang Electric Power Co (Ltd.) and the National Natural Science Foundation of China.

References

[1] P. Boken and G. Callaghan, G, Confronting the challenges of manual journal entries, Protiviti, Alexandria, VA, 1-4.

[2] X. Tan, A. Hammad and P. Fazio, "Automated code compliance checking for building envelope design," J. Comput. Civ. Eng., 10.1061/(ASCE)0887-3801(2010)24:2(203), 203-211.

[3] C. Eastman, J. Lee, Y. Jeong, and J. Lee, "Automatic rule-based checking of building designs," Autom. Constr., 18(8), 1011-1033.

[4] Singapore Building and Construction Authority, "Construction and real estate network: Corenet systems," [http://www.corenet.gov.sg/] (Dec. 15, 2016).

[5] AEC3, "International Code Council," [http://www.aec3.com/en/5/5_013_ICC.htm] (December. 15, 2016).

[6] Fiatech, "Automated code plan checking tool," [http://fiatech.org/active-projects/593-smartcodes%20.html] (July. 16, 2015).

[7] G. Corke, "Solibri model checker V8," AECMagazine: Building information modelling (BIM) for architecture, engineering and construction, [http://aecmag.com/index.php?option=content&task=view&id=527] (May 23, 2016).

[8] Avolve Software Corporation, "Electronic plan review for building and planning departments," [http://www.avolvesoftware.com/index.php/solutions/building-departments/] (Jul. 15, 2011).

[9] Zhong, B. T., Ding, L. Y., Luo, H. B., Zhou, Y., Hu, Y. Z., and Hu, H. M, "Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking," Autom. Constr., 28(2012), 58-70.

[10] Zhang J. and El-Gohary N., "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking," Journal of Computing in Civil Engineering, 30(2), 04015014-1-14.

[11] HanLP, http:// hanlp.linrunsoft.com, (Nov. 15, 2018).

[12] Huijsen and Willem-Olaf, "Controlled Language -An Introduction."

[13] Aerospace and Defence Industries Association of Europe Simplified Technical English Maintanence Group, STE specification, http://www.asd-ste100.org/, (Nov. 15, 2018).

[14] R. Schwitter, K. Kaarel, C. Anne, D. Catherine and H. Glen, "A comparison of three controlled natural languages for OWL 1.1," Proceedings of OWLED 2008, CEUR, vol. 496.

About Author (s):

Wu Bing is currently responsible for the Design Center of the Economic and Technological Research Institute of Zhejiang Electric Power Co., Ltd. She has been doing research on the 3D design technology of power transmission and transformation engineering. As the key member, she participated in the *research* focusing on 3D model standardization of power grid infrastructure and other national grid engineering projects.